Adaptive Generalized Fused-Lasso: Asymptotic Properties and Applications

Vivian Viallon, Sophie Lambert-Lacroix, Holger Höfling, Franck Picard

Adaptive Generalized Fused-Lasso: Asymptotic Properties and Applications

V. Viallon[‡], S. Lambert-Lacroix[†], H. Hoefling[◇] and F. Picard[⋆]

[‡] *Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron.*

[⋆] *LBBE, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne, France*

[◇] *Novartis Pharma, Basel, Switzerland*

[†] *UMR 5525 UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG, Grenoble, F-38041, France*

viallon@math.univ-lyon1.fr sophie.lambert@imag.fr hhoeflin@gmail.com franck.picard@univ-lyon1.fr

Abstract

The Lasso has been widely studied and used in many applications over the last decade. It has also been extended in various directions in particular to ensure asymptotic oracle properties through adaptive weights (Zou, 2006). Another direction has been to incorporate additional knowledge within the penalty to account for some structure among features. Among such strategies the Fused-Lasso (Tibshirani et al., 2005) has recently been extended to penalize differences of coefficients corresponding to features organized along a network, through the Generalized Fused-Lasso. In this work we investigate the theoretical and empirical properties of the Adaptive Generalized Fused-Lasso in the context of Generalized Linear Models, with emphasis on Logistic Regression. More precisely, we establish its asymptotic oracle properties and propose an extensive simulation study to explore its empirical properties. We especially show that it compares favorably with other strategies. We also propose an adaptation of the Relaxed Lasso (Meinshausen, 2007). Finally we present an original application of the Generalized Fused-Lasso to the Joint Modeling framework where the design itself suggests the graph to be used in the penalty; an illustration is provided on road safety data.

[◇] The views and opinions expressed herein are those of the author and do not necessarily reflect the views of Novartis.

Contents

1. Introduction

This paper deals with the general framework of regression that aims at unraveling relationships between a response variable Y and a vector of p covariates or features $\mathbf{x} \in \mathbb{R}^p$. From a practical standpoint, a good comprehension of this relationship is notably useful for prediction matters but also to understand the dynamics of variable Y itself that may be driven by some influential components of $\mathbf{x}$. In Statistics, a natural way to study the relationship between Y and $\mathbf{x}$ is to first assume that this relationship can be correctly approximated by some simple model like Generalized Linear Models (McCullagh and Nelder, 1989) and then estimate the vector of parameters $\boldsymbol{\beta}$ of this model. Estimation is generally performed by maximizing the log-likelihood of the model, yielding Maximum Likelihood Estimators (MLEs). Under mild conditions, these estimators, and in turn, prediction based on these estimators, enjoy good statistical properties (see, *e.g.*, Fahrmeir and Kaufmann (1985)). However, their variance are increasing functions of the number of parameters, which scales as p for the models mentioned above. In situations where only a fraction of the components of $\mathbf{x}$ are expected to be relevant, model selection then becomes desirable: in addition to yield more interpretable models, it reduces the effective number of parameters to be estimated, hence the variance of both estimates and predictions. Another popular approach is to maximize a penalized version of the log-likelihood. Penalties based on ℓ_q norms of parameter $\boldsymbol{\beta}$, $q \geq 0$, lead to shrunk estimates: absolute values of the estimates are biased towards zero. By sacrificing some bias, the variance of the predicted values can be decreased and shrinkage may improve the overall prediction accuracy. Moreover, penalties based on ℓ_q norms of the parameters for $q \leq 1$ encourage sparsity in the vector of parameters, hence model selection. They are methods of choice to return interpretable models that can enjoy good prediction properties.

The most famous example of such a method is the "Least Absolute Shrinkage and Selection Operator" (Lasso) which was originally based on a penalization of the least-squares criterion using the ℓ_1-norm of the parameters (Tibshirani, 1996). If considering a vector of parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, with β_0 the intercept term, the original Lasso penalty is $\text{pen}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$, with λ some tuning parameter. Since then, it has been extended to Generalized Linear Models (Van de Geer, 2008; Friedman et al., 2010). The statistical properties of the Lasso have been extensively investigated, both in the low-dimensional (fixed p) and high-dimensional (diverging p or even $p \geq n$) cases (Knight and Fu, 2000; Fan and Li, 2001; Leng et al., 2006; Wang et al., 2007; Zhao and Yu, 2007; Bunea et al., 2007; Bickel et al., 2009). For instance, under some conditions, the Lasso enjoys sparsistency, i.e. model selection consistency: it selects the right components of vector $\mathbf{x}$ with high probability. However, even in the low-dimensional case, Zou (2006) stated that non trivial conditions on the Gram matrix were necessary to ensure the Lasso sparsistency. This condition is closely related to the irrepresentable condition introduced in Zhao and Yu (2007), or the restricted eigenvalue condition of Bickel et al. (2009). We also refer the reader to Section 6.13 in Bühlmann and Van De Geer (2011) for more details about

these conditions. This has motivated some (inner) modifications of the Lasso itself. Zou (2006) proposed the Adaptive Lasso with penalty $\text{pen}(\boldsymbol{\beta}, \mathbf{w}) = \lambda \sum_{j=1}^{p} |\beta_j|/w_j$ which uses adaptive weights w_j for a differential penalization in the ℓ_1 norm. In the fixed p setting, Zou (2006) further showed that the Adaptive Lasso enjoys an asymptotic oracle property with no particular assumption on the design matrix: as n grows to infinity, it identifies the correct subset model with probability tending to one and estimates of non-zero components perform as well as if the true underlying model were given in advance.

Another way to improve the Lasso is to consider additional terms in the penalty. For instance the Elastic-Net approach proposed by Zou and Hastie (2005) consists in adding a quadratic penalty term to the initial ℓ_1 penalty such that $\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$. The Elastic-Net is particularly desirable in the presence of highly correlated features. Extensions of the Elastic-Net have been introduced to handle structured features. For instance, the Smooth Lasso (S-Lasso) relies on an ℓ_2-fusion penalty such that $\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{j>1} (\beta_j - \beta_{j-1})^2$ (Land and Friedman, 1996) (this is an extension of the Elastic-Net because it can be shown that $\sum_{j>1}(\beta_j - \beta_{j-1})^2 = \boldsymbol{\beta}^T J^T J \boldsymbol{\beta} =: \|\boldsymbol{\beta}\|_J^2$, for some matrix J; see Hebiri and van De Geer (2011)). It is appealing when features can be ordered and their coefficients vary smoothly. Adaptive versions of the Elastic-Net (Ghosh, 2007; Zou and Zhang, 2009) and of the S-Lasso have also been proposed (El Anbari and Mkhadri, 2013). From a theoretical standpoint, these methods return estimates that enjoy sparsistency under weaker conditions than the ones required by the Lasso; see Yuan and Lin (2007); Jia and Yu (2010); Hebiri and van De Geer (2011); El Anbari and Mkhadri (2013).

In fact, the S-Lasso was inspired by the Fused-Lasso of Tibshirani et al. (2005) that has been proposed to enforce similarity between the effects of successive features by using penalty $\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j |\beta_j - \beta_{j-1}|$. Theoretical properties of Fused-Lasso estimates were studied in the Gaussian sequence model (which can be reformulated as a linear regression model with the identity matrix as the design matrix, so that $p = n$ for this model) by Rinaldo (2009) and Qian and Jia (2012). Rinaldo (2009) worked under the assumption that the variance of the noise vanishes as n grows to infinity and established oracle prediction inequalities for a modified version of the Lasso estimate (he called this version adaptive Fused Lasso but no adaptive weight are invoked). In Qian and Jia (2012), the authors work under the assumption that the variance of the noise is constant and show that Fused-Lasso estimates are generally not able to recover the signal pattern. In the more general linear regression setting, Tibshirani et al. (2005) established some asymptotic properties for the Fused-Lasso estimates, in the simpler case of fixed p. More recently, Vaiter et al. (2011) established the sparsistency of the method in the high-dimensional case under a bounded-noise assumption (and of course some assumption on the Gram matrix).

The Fused-Lasso has been generalized by Höfling et al. (2010) to handle complex structure among features effects. The motivation of the Generalized Fused-Lasso is that

when available, *prior* information regarding the structure of features effects should be used to increase the selection and prediction performance. The procedure consists in using an external graph G defined by a set of p vertices V that stands for the p components of vector $\boldsymbol{\beta}^*$, and a set of edges E, such that two connected coefficients are supposed to vary smoothly. This leads to the Generalized Fused-Lasso penalty $\text{pen}(\boldsymbol{\beta}, G) = \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{(j,\ell)\in E} |\beta_j - \beta_\ell|$. The algorithm proposed by Höfling et al. (2010) enables to introduce weights in the penalty term following the idea of the Adaptive Lasso proposed by Zou (2006). They give fast algorithms for solving the Adaptive Generalized Fused Lasso which are based on coordinate-wise optimization but do not present any theoretical results. In this work we propose to investigate the theoretical as well as the empirical properties of the Adaptive Generalized Fused-Lasso. In Section 3 we first prove that the resulting estimator enjoys asymptotic oracle properties, as n goes to infinity and p is fixed, in the context of both linear and logistic regression models. Our asymptotic results constitute a milestone for future extension to the case of diverging p. Using simulation studies we show the empirical benefits of using a ℓ_1-based fusion penalty on support recovery and prediction, compared with other strategies (Section 4). We also show the benefits of using adaptive weights and/or relaxation on Generalized Fused-Lasso estimates. Finally we propose in Section 5 an original application of the Generalized Fused-Lasso to the framework of Joint Modeling, where the design of the study provides the graph to be used in the penalty, and we apply this approach on road-safety data.

2. The Adaptive Generalized Fused-Lasso in GLMs

2.1 Model and loss functions

In this paper, we focus on two particular cases of generalized linear models (McCullagh and Nelder, 1989): the usual linear regression model and the logistic regression model. We mention that theoretical results for other generalized linear models would follow from similar arguments. We decided to focus only on linear and logistic regression models because they are the only ones handled in the algorithms of Höfling et al. (2010), and also for the sake of clarity, since notations can become cumbersome when dealing with the whole class of generalized linear models. For $i = 1, \ldots, n$, let Y_i and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ be the *response* variable and a p-dimensional vector of features (or covariates) respectively. We consider the case of a *fixed design*, *i.e.*, Y_i is a random variable but vectors of features $\mathbf{x}_i$ are assumed to be fixed (non-random). Without loss of generality, we further assume that $\sum_{i=1}^n x_{ij} = 0$. Finally define $\mathbf{z}_i = (1, \mathbf{x}_i^T)^T$.
The linear regression model is very standard, notably when the response variable is continuous. It writes

$$Y_i = \mathbf{z}_i^T \boldsymbol{\beta}^* + \sigma\epsilon_i,$$

where ϵ_i denotes some random noise, and $\sigma > 0$ is fixed and unknown. The vector of coefficients $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ is unknown and has to be estimated. More precisely, $\boldsymbol{\beta}^* =$

$(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T$ where β_0^* denotes the intercept parameter and $\boldsymbol{\beta}_{\setminus 0}^* = (\beta_1^*, \ldots, \beta_p^*)$ corresponds to the regression coefficients pertaining to the covariates. Under this model, estimation is generally performed by ordinary-least squares, which consists in minimizing the squared error loss J_{sq} defined by

$$J_{\text{sq}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{J}_{\text{sq}}(Y_i, \mathbf{z}_i^T \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \mathbf{z}_i^T \boldsymbol{\beta}\right)^2, \tag{1}$$

where function $\mathcal{J}_{\text{sq}} : (u, v) \in \mathbb{R}^2 \mapsto (u - v)^2/2$ is introduced for future use.
When the response variable is binary ($Y_i \in \{0, 1\}$ is often called *label* or *class*), the logistic model is the most standard one. Introducing the *logit* function, $\text{logit}(x) = \log(x/(1 - x))$, the class probability, or equivalently the expectation of Y_i, is $\pi_i = \mathbb{P}(Y_i = 1) = \text{logit}^{-1}(\mathbf{z}_i^T \boldsymbol{\beta}^*) = 1/(1 + \exp(-\mathbf{z}_i^T \boldsymbol{\beta}^*))$, where logit^{-1} is the reciprocal of the logit function. Again, the vector $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ is unknown and has to be estimated. Estimation is usually performed by maximizing the log-likelihood of the model, which is equivalent to minimizing the logistic loss function J_{lo}, given by

$$J_{\text{lo}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{J}_{\text{lo}}(Y_i, \mathbf{z}_i^T \boldsymbol{\beta}) = -\sum_{i=1}^{n} \left\{ Y_i \mathbf{z}_i^T \boldsymbol{\beta} - \log\left(1 + \exp(\mathbf{z}_i^T \boldsymbol{\beta})\right) \right\}, \tag{2}$$

where function $\mathcal{J}_{\text{lo}} : (u, v) \in \mathbb{R}^2 \mapsto -uv + \log(1 + \exp(v))$ is introduced for future use.

2.2 The Adaptive Generalized Fused-Lasso penalty

We focus on ℓ_1-based Fused penalties inspired from the Fused-Lasso of Tibshirani et al. (2005). In this framework features $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ correspond to p successive positions, and to account for sparsity in terms of successive differences, the original Fused penalty is defined by:

$$\text{pen}(\boldsymbol{\beta}) = \lambda_n^{(1)} \sum_{j=1}^{p} |\beta_j| + \lambda_n^{(2)} \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|,$$

and depends on two tuning parameters $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$. Then this Fused framework has been generalized by Höfling et al. (2010) to the case of networks of features. This Generalized Fused penalty is of great interest when connected features along some network may have similar effects on the response. Consider a graph $G = (V, E)$, with node set V that corresponds to the indices of coefficients in $\boldsymbol{\beta}_{\setminus 0}$ (*i.e.* $V = \{1, ...p\}$), and edge set E that corresponds to pairs of connected coefficients indices (j, ℓ) with $j > \ell$. The graph G that is used in the penalty corresponds to some *prior* knowledge given by an expert, and hence is fixed. The Generalized Fused-Lasso penalty consists in penalizing all coefficient differences for which an edge exists in G:

$$\text{pen}(\boldsymbol{\beta}; G) = \lambda_n^{(1)} \sum_{j \in V} |\beta_j| + \lambda_n^{(2)} \sum_{(j,\ell) \in E} |\beta_j - \beta_\ell|. \tag{3}$$

The Fused-Lasso penalty of Tibshirani et al. (2005) corresponds to a Generalized Fused-Lasso penality based on a chain graph, where $(j, \ell) \in E$ if and only if $\ell = j - 1$ (see Figure 1 for a simple illustration).

Adaptive weights can further be introduced following the idea of the Adaptive Lasso proposed by Zou (2006). This results in the *Adaptive* Generalized Fused-Lasso penalty:

$$\text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}) = \lambda_n^{(1)} \sum_{j \in V} w_j^{(1)} |\beta_j| + \lambda_n^{(2)} \sum_{(j,\ell) \in E} w_{j\ell}^{(2)} |\beta_j - \beta_\ell|.$$

where $w_j^{(1)}$ and $w_{j\ell}^{(2)}$ are weights vectors. In this paper (where p is fixed) they are based on initial Maximum-Likelihood (ML) estimates $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$. More precisely, they are set to $w_j^{(1)} = |\widetilde{\beta}_j|^{-\gamma}$ and $w_{j\ell}^{(2)} = |\widetilde{\beta}_j - \widetilde{\beta}_\ell|^{-\gamma}$, where γ is some fixed positive constant. The rationale is to penalize more heavily coefficients (or differences of coefficients) when their ML estimates are small: the higher γ, the more trust is put into the ML estimates. A typical value for γ is 1 (value that is used in our simulations and applications). The Adaptive Generalized Fused criterion Q is then simply defined, for given graph G and weights $\mathbf{w}$, as

$$Q(\boldsymbol{\beta}) = J(\boldsymbol{\beta}) + \text{pen}_{\text{Ada}}(\boldsymbol{\beta}; G, \mathbf{w}), \tag{4}$$

where the loss function is $J = J_{\text{lo}}$ or $J = J_{\text{sq}}$ depending on the working model (see equations (1) and (2) above).

2.3 Application of the Generalized Fused penalty to Joint Modeling

Interestingly, another source of knowledge concerning features structure can be provided by the design of the study itself. In this work, we propose an original application of the Generalized Fused-Lasso framework to the case of joint estimation of multiple sparse regression models. The joint modeling framework described here has some connections with seemingly unrelated regression problems and multi-task learning (see Huang et al. (2012) and the references therein). More specifically we consider the very common case of data collected from distinct *strata*, which often arises in epidemiology where each stratum can be defined by crossing gender, age and ethnicity for instance. The design is structured according to a known (and fixed) categorical vector $(\mathcal{C}_1, \ldots, \mathcal{C}_n)$ taking values in $\{1, \ldots, C\}$, with $C \geq 1$ the total number of strata. Let n_c be the number of observations falling into stratum c (so that $n = \sum_c n_c$). Without loss of generality, we further assume that the n observations are ordered so that the first n_1 observations correspond to stratum 1 (that is $\mathcal{C}_i = 1$ for $i = 1, \ldots, n_1$), the next n_2 observations correspond to stratum 2, and so forth. In the case of a linear regression model, we would have

$$Y_i = \mathbf{z}_i^T \boldsymbol{\beta}^*_{\mathcal{C}_i} + \sigma \epsilon_i \qquad \text{for } i = 1, \ldots, n,$$

where $\boldsymbol{\beta}^*_c$, $c = 1, \ldots, C$, denotes the vector of parameter for stratum c. The purpose of the analysis is to determine whether the distribution of the response varies across strata,

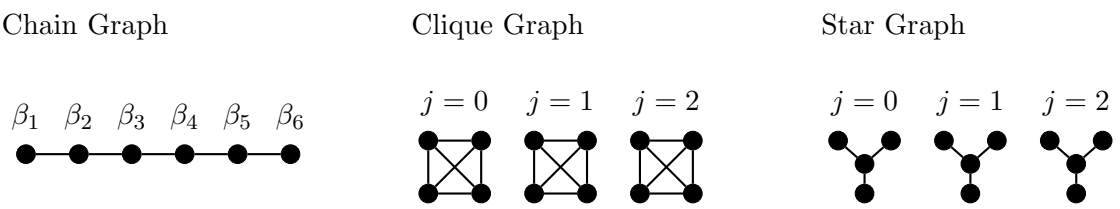


Figure 1: Typical examples of graphs used in the Generalized Fused-Lasso penalty. ***(Left)*** The chain graph is typically used when covariates are naturally ordered as is the case for CGH array. This example illustrates the situation where $p = 6$. ***(Middle)*** The clique graph is typically used in the joint modeling context ($p = 2$ and $C = 4$ in this example). ***(Right)*** The star graph is typically used in the joint modeling context, when one stratum serves as the reference ($p = 2$ and $C = 4$ in this example).

i.e. to detect which components of $\boldsymbol{\beta}_c^*$ do vary with c. Constructing independent (possibly sparse) models for each stratum would not take advantage of the common structure, while constructing a single model for the whole data set would mask the differences. Alternatively, the Generalized Fused-Lasso can be used to couple estimations obtained from each stratum, encouraging them to share a common structure. More precisely, we propose the following penalty:

$$\sum_{c=1}^{C} \left\{ \lambda_n^{(1)} \sum_{j=1}^{p} w_j^{(1)} |\beta_{c,j}| \right\} + \lambda_n^{(2)} \sum_{j=0}^{p+1} \sum_{c_1 > c_2} w_{c_1,c_2,j}^{(2)} |\beta_{c_1,j} - \beta_{c_2,j}|,$$

where $w_j^{(1)}$ and $w_{c_1,c_2,j}^{(2)}$ are appropriate adaptive weights. Parameter $\lambda_n^{(2)}$ governs the amount of shrinkage for differences between strata: if null, this penalty resumes to C independent Lasso penalties. If positive, the Fused part of the penalty encourages coefficients $\beta_{c_1,j}$ and $\beta_{c_2,j}$ to be at least close to each other (that is, the j^{th} coefficient in strata c_1 and c_2 respectively). In Appendix we give more details on particular aspects of Joint modeling: we show that the general context described here induces a clique graph in the penalty, while the case of designs with a control stratum induces a star graph instead (see Figure 1 for a simple illustration).

3. Theoretical results

In this section, we study the asymptotic properties of the Adaptive Generalized Fused-Lasso estimator in both the Gaussian and logistic regression contexts. More precisely, we assume that p is fixed and let n grow to infinity. The case where the dimension p may grow as the sample size n increases is much more tricky and is left for future research.

As mentioned above, our results readily extend to other generalized linear models, under the same assumptions as **AL1** and **AL2** below, and using the same arguments as the ones used in our proofs (see Sections 7.2 and 7.4 in the Appendix).
Before stating our results some notations and assumptions need to be introduced. Let $\mathcal{A} = \{1 \leq j \leq p,\, \beta_j^* \neq 0\}$ be the *support* of $\boldsymbol{\beta}^*_{\backslash 0}$ and $p_0 = |\mathcal{A}|$ its cardinality (*i.e.*, the true vector of regression parameters $\boldsymbol{\beta}^*_{\backslash 0}$ is assumed to have p_0 non-zero elements). Further consider the set

$$\mathcal{B} = \{(j, \ell) \in E,\, \beta_j^* \neq 0 \text{ and } \beta_j^* = \beta_\ell^*\}.$$

For future use, note that $\mathcal{B} \subset \mathcal{A} \times \mathcal{A}$. For any integer k, we denote by $\mathbf{0}_k$, $\mathbf{1}_k$ and $\mathbf{I}_k$ the vector of 0's and 1's in $\mathbb{R}^k$ and the identity matrix of size k. For any subset $\mathcal{S} \subseteq \{1, \ldots, p\}$, we denote by $\bar{\mathcal{S}}$ its complement in $\{1, \ldots, p\}$, and for any vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ we further denote by $\boldsymbol{\beta}_{\mathcal{S}}$ the vector given by the coordinates of $\boldsymbol{\beta}$ the index of which are in $\{0\} \cup \mathcal{S}$. Moreover, we denote by $\|\cdot\|$ the usual Euclidian norm. For any $x \in \mathbb{R}$, we define the function $sign(x)$ which equals $+1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. We also introduce $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{Y} = [Y_1, \ldots, Y_n]^T$. Finally, for any (possibly random) event Ω, we denote by $\mathbb{I}(\Omega)$ the indicator function which equals 1 if Ω is true and 0 otherwise.
In the Gaussian context, we work under the following assumptions.

- **A1** Variables $\epsilon_1, \ldots, \epsilon_n$ are independent identically distributed random variables with mean 0 and variance 1;
- **A2** Matrix $\mathbf{Z}^T\mathbf{Z}/(n\sigma^2)$ converges to $\mathbf{C}$ where $\mathbf{C}$ is a positive definite matrix.

In the logistic regression context we denote by $\mathcal{I}(\boldsymbol{\beta})$ the empirical Fisher's matrix of size $(p+1) \times (p+1)$. For future use, observe that $\mathcal{I}(\boldsymbol{\beta}^*) = \mathbf{Z}^T \boldsymbol{D} \mathbf{Z}$, where $\boldsymbol{D}$ denotes the $n \times n$ diagonal matrix with i-th diagonal element given by $\pi_i(1 - \pi_i)$. For any $\delta \geq 0$, we further denote by $N_n(\delta)$ the neighborhood of $\boldsymbol{\beta}^*$ defined by

$$N_n(\delta) = \left\{ \boldsymbol{\beta} \in G / \left\| \left[\mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}} \right]^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\| \leq \delta \right\}.$$

We will work under the following conditions:

- **AL1** $\mathcal{I}(\boldsymbol{\beta}^*)/n$ converges to $\mathbf{C}$ where $\mathbf{C}$ is a positive definite $(p+1) \times (p+1)$ matrix;
- **AL2** As n goes to ∞,

$$\max_{\boldsymbol{\beta} \in N_n(\delta)} \left\| \mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}} \mathcal{I}(\boldsymbol{\beta}^*) \left[\mathcal{I}(\boldsymbol{\beta})^{-\frac{1}{2}} \right]^T - \mathbf{I}_{p+1} \right\| \to 0.$$

Assumptions **AL1** and **AL2** are standard when working under generalized linear models (McCullagh and Nelder, 1989). Assumption **AL1** is similar to **A2**, which was used by,

e.g., Tibshirani et al. (2005) in their study of the Fused-Lasso in the Gaussian context. Let us remark that under **AL1** or **A2**, criterion (4) corresponds, for n large enough, to a strong convex optimization problem, and thus is not concerned by the issue of multiple local minima.

We first state an asymptotic result for logistic regression models in the non adaptive case, which is similar to Theorem 1 of Tibshirani et al. (2005), established in the Gaussian case for the Generalized Fused-Lasso based on chain-graphs. The proof is given in the Appendix (Section 7.2).

Theorem 1 *Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of criterion Q defined in (4) with $J = J_{\rm lo}$ or $J = J_{\rm sq}$ in the non adaptive case (that is with $w_j^{(1)} = 1$ and $w_{j\ell}^{(2)} = 1$ for all j, ℓ). If $\lambda_n^{(m)}/\sqrt{n} \to \lambda_0^{(m)} \geq 0$ ($m = 1, 2$), then under assumptions* **A1-2** *for the Gaussian case and* **AL1-2** *for the logistic case,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \to_d \arg\min(\mathcal{V}),$$

where $\mathcal{V}$ is the function defined, for $\mathbf{u} = (u_0, \ldots, u_p) \in \mathbb{R}^{p+1}$, as

$$\begin{aligned}\mathcal{V}(\mathbf{u}) &= \mathbf{u}^T\mathbf{W} + \frac{a}{2}\mathbf{u}^T\mathbf{C}\mathbf{u} + \lambda_0^{(1)}\sum_{j=1}^{p}\{u_j sign(\beta_j^*)\mathbb{I}(\beta_j^* \neq 0) + |u_j|\mathbb{I}(\beta_j^* = 0)\} \\ &\quad + \lambda_0^{(2)}\sum_{(j,\ell)\in E}\{(u_j - u_\ell)sign(\beta_j^* - \beta_\ell^*)\mathbb{I}(\beta_j^* \neq \beta_\ell^*) + |u_j - u_\ell|\mathbb{I}(\beta_j^* = \beta_\ell^*)\}.\end{aligned}$$

Above, $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}_{p+1}, \mathbf{C})$ distribution and $a = 1$ in the logistic case and $a = \sigma^2$ in the linear case.

This result establishes the root-n consistency of non-adaptive Generalized Fused-Lasso estimates. However, this result also implies that when $\lambda_n^{(m)} = O(\sqrt{n})$, for $m = 1, 2$, the support of $\boldsymbol{\beta}^*$ can not be recovered with high probability by non-adaptive Fused-Lasso estimates, as stated in the following proposition whose proof is given in the Appendix (see Section 7.3).

Proposition 2 *Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of criterion Q defined in (4) with $J = J_{\rm lo}$ or $J = J_{\rm sq}$ in the non adaptive case (that is with $w_j^{(1)} = 1$ and $w_{j\ell}^{(2)} = 1$ for all j, ℓ). Further set $\widetilde{\mathcal{A}}_n = \{1 \leq j \leq p,\, \widehat{\beta}_j \neq 0\}$. If $\lambda_n^{(m)}/\sqrt{n} \to \lambda_0^{(m)} \geq 0$ ($m = 1, 2$), then under assumptions* **A1-2** *for the Gaussian case and* **AL1-2** *for the logistic case,*

$$\limsup_n \mathbb{P}(\widetilde{\mathcal{A}}_n = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

We now show that for appropriate choices of $\lambda_n^{(m)} = O(\sqrt{n})$ for $m = 1, 2$, the Adaptive Generalized Fused-Lasso estimator $\widehat{\boldsymbol{\beta}}^{ad}$, defined as the minimizer of criterion Q in (4), enjoys asymptotic oracle properties in both the Gaussian and logistic regression contexts, contrasting with its non-adaptive counterpart. We introduce $\mathcal{A}_n = \{1 \leq j \leq p,\, \widehat{\beta}_j^{ad} \neq 0\}$ and

$$\mathcal{B}_n = \{(j,\ell) \in E,\, \widehat{\beta}_j^{ad} \neq 0 \text{ and } \widehat{\beta}_j^{ad} = \widehat{\beta}_\ell^{ad}\}.$$

Some more notations are needed before stating our result: in particular, the number s_0 of distinct non-zero values in $\boldsymbol{\beta}^*_{\setminus 0}$ "supported" by G needs to be precisely defined. The edge set E can be decomposed as $E = E_{\mathcal{A}} \cup E_{\bar{\mathcal{A}}}$ where $E_{\mathcal{A}} = \{(j,\ell) \in E : \beta_j^* \beta_\ell^* \neq 0\}$ and $E_{\bar{\mathcal{A}}} = \{(j,\ell) \in E : \beta_j^* \beta_\ell^* = 0\}$ correspond to the set of edges made of vertices in $\mathcal{A}$ only and the set of edges made of at least one vertex in $\bar{\mathcal{A}}$ respectively. Subset $E_{\mathcal{A}}$ can further be decomposed as $E_{\mathcal{A}} = E_{\mathcal{A}}^{\neq}(\boldsymbol{\beta}^*) \cup E_{\mathcal{A}}^{=}(\boldsymbol{\beta}^*)$ with $E_{\mathcal{A}}^{=}(\boldsymbol{\beta}^*) = \{(j,\ell) \in E_{\mathcal{A}} : \beta_j^* = \beta_\ell^*\} = \mathcal{B}$ and $E_{\mathcal{A}}^{\neq}(\boldsymbol{\beta}^*) = \{(j,\ell) \in E_{\mathcal{A}} : \beta_j^* \neq \beta_\ell^*\}$. Then consider the graph $G_{\mathcal{B}} = (\mathcal{A}, \mathcal{B})$ and denote by s_0 the number of its connected components (e.g., in the particular case where G is a chain graph, s_0 is the number of segments consisting of non-zero and equal coefficients). Observe that $d_0 \leq s_0 \leq p_0$, where $p_0 = |\mathcal{A}|$ is the number of non-zero components in $\boldsymbol{\beta}^*_{\setminus 0}$ and d_0 is the number of *distinct* non-zero values in $\boldsymbol{\beta}^*_{\setminus 0}$. We actually have $s_0 = p_0$ if and only if $(\beta_j^* = \beta_\ell^* \neq 0 \Rightarrow (j,\ell) \notin E)$. On the other hand, $s_0 = d_0$ if and only if for all (j,ℓ) such that $\beta_j^* = \beta_\ell^*$, j and ℓ belong to the same connected component of $G_{\mathcal{B}}$. Now denote by $\mathcal{A}_1, \ldots \mathcal{A}_{s_0}$ the sets of vertices of each connected components of $G_{\mathcal{B}}$, with $\mathcal{A} = \bigcup_{s=1}^{s_0} \mathcal{A}_s$, and set $j_s = \min\{\mathcal{A}_s\}$ for $s = 1, \ldots, s_0$. Now we can define $\boldsymbol{\beta}^*_{\mathcal{B}} = (\beta_0^*, \beta_{j_1}^*, \ldots, \beta_{j_{s_0}}^*)^T$, which is composed by the intercept and the s_0 distinct non-zero values of $\boldsymbol{\beta}^*_{\setminus 0}$ supported by G; we further set $\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}} = (\hat{\beta}_0^{ad}, \hat{\beta}_{j_1}^{ad}, \ldots, \hat{\beta}_{j_{s_0}}^{ad})^T$ its estimate. Now denote by $\mathbf{X}_{\mathcal{B}}$ the matrix of size $n \times s_0$, whose s-th column, $1 \leq s \leq s_0$, is $X_{\mathcal{B}_s} = \sum_{j \in \mathcal{A}_s} X_j$, where X_j is the j-th column of $\mathbf{X}$. Finally set $\mathbf{Z}_{\mathcal{B}} = (\mathbf{1}_n, \mathbf{X}_{\mathcal{B}})$ and denote by $\mathbf{C}_{\mathcal{B}}$ the $(s_0+1) \times (s_0+1)$ positive definite matrix that is defined as the limit, as $n \to \infty$, of (i) $({\mathbf{Z}_{\mathcal{B}}}^T \mathbf{Z}_{\mathcal{B}})/(n\sigma^2)$ in the linear case and (ii) $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}})/n$ in the logistic case, where $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}})$ denotes the Information matrix of the model induced by $\mathcal{B}$, that is $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{B}}) = \mathbf{Z}_{\mathcal{B}}^T \boldsymbol{D} \mathbf{Z}_{\mathcal{B}}$. We have now all the ingredients to state our main result.

Theorem 3 *If $\lambda_n^{(m)}/\sqrt{n} \to 0$ and $\lambda_n^{(m)} n^{(\gamma-1)/2} \to \infty$, $(m = 1, 2)$, then, under assumptions* **A1-2** *for the Gaussian case and* **AL1-2** *for the logistic case, the Adaptive Generalized Fused-Lasso estimator satisfies the following properties:*

1. *Consistency in variable selection:* $\mathbb{P}[\mathcal{A}_n = \mathcal{A}] \to 1$ *and* $\mathbb{P}[\mathcal{B}_n = \mathcal{B}] \to 1$ *as* $n \to +\infty$.
2. *Asymptotic normality:* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{ad}_{\mathcal{B}} - \boldsymbol{\beta}^*_{\mathcal{B}}\right) \longrightarrow_d \mathcal{N}\left(\mathbf{0}_{s_0+1}, {\mathbf{C}_{\mathcal{B}}}^{-1}\right)$.

The proof of Theorem 3 is provided in the Appendix (see Section 7.4).

4. Simulation study

We perform an extensive simulation study to explore the performance of the Adaptive Generalized Fused-Lasso penalties in Logistic Regression. Our objectives are:

1. to illustrate our theoretical results using designs inspired from Zou (2006) with $p = 4$ on Adaptive Fused penalties based on chain graphs,
2. to study the influence of the network provided to the Generalized Fused penalty on the performance in terms of selection, estimation and prediction,
3. to show the interest of using the Generalized Fused-Lasso in the context of joint modeling.

We mention that simulations were also performed in the context of linear regression, which lead to similar conclusions (not shown).

4.1 Simulations Setting

Setting $(\mathbf{Y}, \mathbf{X})$. Our theoretical results being stated in the fixed design setting, we first define N as the maximal sample size considered in a given scenario (for instance if samples of size $n = 24, 120, 240$ and 1200 were considered then $N = 1200$), and we generate N *i.i.d.* predictors $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, N$ from a $\mathcal{N}(\mathbf{0}_p, \mathbf{C}_{\setminus 0 \setminus 0})$ distribution, where $\mathbf{C}_{\setminus 0 \setminus 0}$ is some given covariance matrix of size $p \times p$. Then, given $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\boldsymbol{\beta}^*$, the vector of true coefficients (see the paragraph below for the choice of $\boldsymbol{\beta}^*$), the n-vector of labels is generated according to a Bernoulli distribution such that $Y_i \sim \mathcal{B}(\pi_i)$, with π_i defined as in Section 2. Fifty such vectors of labels are generated and results presented below correspond to averages over these 50 replicates (confidence intervals based on the central limit theorem also appear on our figures).

Choosing $\boldsymbol{\beta}^*$. In all our experiments, the true intercept term β_0^* is set to 0. Then the remaining components of vector $\boldsymbol{\beta}^*$ are chosen to achieve a given *Signal to Noise Ratio* (SNR), which is defined as $\mathrm{SNR}^2(\mathbf{Z}, \boldsymbol{\beta}^*) = \|\mathbf{Z}\boldsymbol{\beta}^*\|^2/(n\sigma^2)$ in linear regression. In Generalized Linear Models the link function should be accounted for in the definition of the covariates effects, which motivates a likelihood-based generalization of the SNR such that:

$$\mathrm{SNR}(\mathbf{Z}, \boldsymbol{\beta}^*) = \frac{\sum_{i=1}^n \mathbb{E}\left(\mathcal{J}_{\mathrm{lo}}(Y_i, \bar{\beta}) - \mathcal{J}_{\mathrm{lo}}(Y_i, \mathbf{z}_i^T\boldsymbol{\beta}^*)\right)}{\sum_{i=1}^n \mathbb{E}\mathcal{J}_{\mathrm{lo}}(Y_i, \bar{\beta})},$$

with $\mathcal{J}_{\mathrm{lo}}(Y_i, \bar{\beta})$ the log-likelihood of the null model. This generalization reduces to a R^2 in linear regression (Heinzl and Mittlböck, 2003).

Choosing $G = (V, E)$. For a given true vector of coefficients $\boldsymbol{\beta}^*$ the best graph is the one whose edges correspond to equal coefficients only ($(j, \ell) \in E$, with $j > \ell$, if and only if $\beta_j^* = \beta_\ell^*$), and the worst graph is the one whose edges correspond to coefficients

	$\{\beta_j^* = \beta_\ell^*\}$	$\{\beta_j^* \neq \beta_\ell^*\}$
$(j,\ell) \in E$	TN differences	FN differences
$(j,\ell) \notin E$	FP differences	TP differences
Criteria	$\texttt{Spe}_G = \texttt{TN}/(\texttt{TN}+\texttt{FP})$	$\texttt{Sens}_G = \texttt{TP}/(\texttt{TP}+\texttt{FN})$

Table 1: Measures of the suitability of a particular graph $G = (V, E)$, for a given true vector of coefficients $\boldsymbol{\beta}^*$: sensitivity-like and specificity-like criteria. FP for False Positive, TP for True Positive, FN for False Negative and TN for True Negative.

of different values only. In the latter case, the graph in the penalty provides a false information regarding pairs of equal coefficients. Between these two extreme cases, there exists a continuum of discrepancies between informations supported by E and the true differences in $\boldsymbol{\beta}^*$: when choosing a particular graph, the statistician actually tries to guess the true structure of $\boldsymbol{\beta}^*$ and, of course, the better the guess the better the expected results. The suitability of graph $G = (V, E)$, *i.e.*, the accuracy of the guess, can be measured by sensitivity-like and specificity-like criteria. We denote by $\texttt{Sens}_G$, the proportion of null differences that are supported by the graph among true null differences, and $\texttt{Spe}_G$ the proportion of non-null differences that are not supported by the graph among true non-null differences (see Table 1). The highest these criteria the most accurate the guess, and the most suitable is the graph for the problem at hand. In some experiments below, we make this suitability vary to investigate its impact on selection, estimation and prediction performance of Adaptive Generalized Fused-Lasso estimates.

4.2 Implementation

The Adaptive Generalized Fused Lasso was solved with the algorithms described in Höfling et al. (2010); the corresponding R package will be made available soon.

The Relaxed Adaptive Generalized Fused-Lasso. The Relaxed Lasso of Meinshausen (2007) has been shown to lead to less biased estimates than the crude Lasso: we therefore implemented Relaxed versions for the Adaptive Generalized Fused-Lasso (as well as for each of the competing method presented below), setting coefficient ϕ of Meinshausen (2007) to 0. These methods return unshrunk (unbiased) estimates of non-null coefficients that are obtained by a 2-step approach. The principle of the Relaxed Adaptive Generalized Fused-Lasso is as follows. For any given pair (λ_1, λ_2), we first run the Adaptive Generalized Fused-Lasso method, and obtain a vector $\widehat{\boldsymbol{\beta}}^{ad}_{\lambda_1,\lambda_2}$. Unshrunk estimates $\widehat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}$ are then obtained by standard (*i.e.*, unpenalized) maximum likelihood computed under a constraint induced by the sparsity and structure of $\widehat{\boldsymbol{\beta}}^{ad}_{\lambda_1,\lambda_2}$. More precisely, adopting notations similar to those used in Rinaldo (2009), there exists a (possibly trivial) partition $\{P_1, \ldots, P_{\hat{j}}\}$ of $\{1, \ldots, p\}$ such that $\widehat{\boldsymbol{\beta}}^{ad}_{\lambda_1,\lambda_2} = \sum_{j=1}^{\hat{j}} \hat{\nu}_j \mathbb{I}_{P_j}$, where $\mathbb{I}_{P_j}$ is the p-dimensional

vector whose kth coordinate is 1 if $k \in P_j$ and 0 otherwise. Further introduce $\hat{s}_0 =$ the number of distinct non-null coefficients of $\widehat{\boldsymbol{\beta}}^{ad}_{\lambda_1,\lambda_2}$. Now, denoting by X_k the k-th column of matrix $\mathbf{X}$, we set $\widetilde{\mathbf{Z}} = (\mathbf{1}_n, \widetilde{\mathbf{X}})$ where $\widetilde{\mathbf{X}}$ is the $[n \times \hat{s}_0]$ matrix with column j defined by $\widetilde{X}_j = \sum_{k \in P_j} X_k$. A vector of unshrunk estimates for the distinct non-null coefficients of $\widehat{\boldsymbol{\beta}}^{ad}_{\lambda_1,\lambda_2}$ can then be obtained as

$$\widetilde{\widehat{\boldsymbol{\beta}}}_{\lambda_1,\lambda_2} = \arg\max_{\boldsymbol{\beta}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left\{ \mathrm{Y_i} \widetilde{\mathbf{z}}_{\mathrm{i}} \boldsymbol{\beta} - \log\left(1 + \exp(\widetilde{\mathbf{z}}_{\mathrm{i}} \boldsymbol{\beta})\right) \right\},$$

where $\widetilde{\mathbf{z}}_i$ is the i-th line of matrix $\widetilde{\mathbf{Z}}$. The vector $\widehat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2} \in \mathbb{R}^{p+1}$ is then easily derived from $\widetilde{\widehat{\boldsymbol{\beta}}}_{\lambda_1,\lambda_2}$. This vector is the one returned by our Relaxed Adaptive Generalized Fused-Lasso.

Selection of the tuning parameters. The Fused-Lasso methods (as well as the competing methods presented below) involve some tuning parameters. Given a (possibly 2-d) grid of potential values for these tuning parameters, two main approaches exist to select the optimal one(s): cross-validation and Bayesian Information Criterion (BIC). In the context considered in this paper (large n and small p), both approaches generally lead to comparable results, the one based on the BIC being generally faster. Therefore, we only consider this latter approach here. For instance, parameters (λ_1, λ_2) of the Relaxed Adaptive Generalized Fused-Lasso method are selected as the minimizers, on a predefined 2-d grid, of the following criterion

$$\mathrm{BIC}(\lambda_1, \lambda_2) = 2J_{\mathrm{lo}}(\widehat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}) + \log(n)\mathrm{df}_{\lambda_1,\lambda_2},$$

where $\mathrm{df}_{\lambda_1,\lambda_2}$ is set to the number of distinct non-null coefficients of $\widehat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}$.

4.3 Competing methods

We compete the Adaptive Generalized Fused-Lasso with the Lasso and two other methods: a version of the Group Lasso with an extra ℓ_1-norm penalty (see Section 7.5 in the Appendix for more details) and a recent method proposed by Sun and Wang (2012) called the Generalized Elastic Net. Given a graph $G = (V, E)$, this method relies on the following penalty

$$\mathrm{pen}(\boldsymbol{\beta}, G) = \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \sum_{(j,\ell) \in E} (\beta_j - \beta_\ell)^2 \right],$$

and is implemented in the `pclogit` `R`-package. The authors interpret this penalty as a graph-based Elastic Net penalty. This penalty has not been extended to handle adaptive weights yet and does not propose two independent tuning parameters.

An important result of our numerous experiments is that the performance of non-fused based penalties (*i.e.* the Lasso, the Generalized Elastic-Net, the Group Lasso and their variants) are all comparable to the performance of the Relaxed Adaptive Lasso (in most cases they are actually indistinguishable). Concerning the Generalized Elastic-Net this illustrates the poor influence of the (graph-based) ℓ_2 norm part of the penalty. This could be explained by *i*) the lack of selection property of the ℓ_2 norm that does not encourage differences to be zero, *ii*) the current implementation of the method: the `pclogit` package uses dependent tuning parameters which may lack of flexibility. It is worth mentioning that the GE-Net is closely related to the Smooth Lasso of Hebiri and van De Geer (2011), whose implentation is available in the linear case only (hence not used in the following simulations that focus on the Logistic case). However we report that when compared to the Fused Lasso in the linear case (not shown), the Smooth Lasso shows better performance than the Lasso in ultra-correlated cases only. As for the Group Lasso, its penalty is composed by the Lasso penalty plus the Group Lasso penalty: our results suggest that this latter penalty may not be appropriate when considering networks of features since it does not improve upon the "pure" Lasso. Consequently results pertaining to the Generalized Elastic-Net and to the Group Lasso are in general not displayed in the sequel for a better legibility of our results (they are displayed in our first experiments for illustration only).

Thus the simulation study focuses on ℓ_1-based Fused penalties only, and we explore the properties of "raw" Generalized Fused-Lasso (referred to as Fused on our figures), Adaptive Generalized Fused-Lasso (Ada-Fused), Relaxed Generalized Fused-Lasso (Relaxed-Fused), and Relaxed Adaptive Generalized Fused-Lasso (Relaxed-Ada-Fused). The comparisons are made using the Relaxed Adaptive Lasso (Relaxed-Ada-Lasso) as a reference that does not account for any network structure among features.

4.4 Evaluation criteria

Accuracy on support recovery. When comparing model selection algorithms, it is common to use criteria originally devoted to the evaluation of binary classification algorithms. For the evaluation of support recovery for instance, each component β_j^* is either zero or non-zero, as is its estimate. Set as before $\mathcal{A} = \{1 \leq j \leq p, \beta_j^* \neq 0\}$ and, for any estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, $\mathcal{A}_n = \{1 \leq j \leq p, \hat{\beta}_j \neq 0\}$. A true positive is then defined as a node that belongs to $\mathcal{A} \cap \mathcal{A}_n$, (i.e. the set of true positives is $\{1 \leq j \leq p,\ \beta_j^* \neq 0, \widehat{\beta}_j \neq 0\}$). Similarly, a true negative is defined as a node that belongs to $\bar{\mathcal{A}} \cap \bar{\mathcal{A}}_n$, (i.e. the set of true negatives is $\{1 \leq j \leq p,\ \beta_j^* = 0, \widehat{\beta}_j = 0\}$). The accuracy on support recovery, denoted by `Acc.A`, is then obtained by adding these two terms and dividing the result by p.

Accuracy on pairs of coefficients. In the Fused framework, we also need to evaluate the performance regarding the classification of pairs of coefficients. As an illustration, in the particular case of the chain-graph penalty (corresponding to consecutive coefficients),

Indicator	support recovery	equality among connected coefficients
TPR	$\lvert\mathcal{A}\cap\mathcal{A}_n\rvert/\lvert\mathcal{A}\rvert$	$\lvert\mathcal{E}^{\neq}(\beta^*)\cap\mathcal{E}^{\neq}(\widehat{\beta})\rvert/\lvert\mathcal{E}^{\neq}(\beta^*)\rvert$
FPR	$\lvert\bar{\mathcal{A}}\cap\mathcal{A}_n\rvert/\lvert\bar{\mathcal{A}}\rvert$	$\lvert\mathcal{E}^{=}(\beta^*)\cap\mathcal{E}^{\neq}(\widehat{\beta})\rvert/\lvert\mathcal{E}^{=}(\beta^*)\rvert$
Acc	$(\lvert\bar{\mathcal{A}}\cap\bar{\mathcal{A}}_n\rvert+\lvert\mathcal{A}\cap\mathcal{A}_n\rvert)/p$	$(\lvert\mathcal{E}^{=}(\beta^*)\cap\mathcal{E}^{=}(\widehat{\beta})\rvert+\lvert\mathcal{E}^{\neq}(\beta^*)\cap\mathcal{E}^{\neq}(\widehat{\beta})\rvert)/\lvert\mathcal{E}\rvert$

Table 2: Evaluation criteria for model selection.

this classification of pairs reduces to the classification of zero and non-zero elements of the vectors of successive differences. Its evaluation enables to assess the capacity of the method to detect equal consecutive coefficients. In the context of Joint Modeling this evaluation enables to assess the method capacity to detect heterogeneity and homogeneity across strata.

Introduce for any vector $\boldsymbol{\beta}\in\mathbb{R}^{p+1}$ and any subset of edges $\mathcal{E}\subseteq E$, the subsets $\mathcal{E}^{=}(\boldsymbol{\beta}):=\{(j,\ell)\in\mathcal{E},\beta_j=\beta_\ell\}$ and $\mathcal{E}^{\neq}(\boldsymbol{\beta}):=\{(j,\ell)\in\mathcal{E},\beta_j\neq\beta_\ell\}$ of edges in $\mathcal{E}$ corresponding to pairs of equal and non-equal components in vector $\boldsymbol{\beta}$ respectively. Considering pairs, a true positive is an edge that belongs to $\mathcal{E}^{\neq}(\boldsymbol{\beta}^*)\cap\mathcal{E}^{\neq}(\widehat{\boldsymbol{\beta}})$; definitions of true negatives and accuracy for pairs of components follows similarly (see Table 2). Regarding the choice of $\mathcal{E}$, we consider the particular choices $\mathcal{E}=E$ (no restriction), $\mathcal{E}=E_{\mathcal{A}}=\{(j,\ell)\in E:\beta_j^*\beta_\ell^*\neq 0\}$ (restriction to edges in the graph corresponding to pairs of non-zero true coefficients) and $\mathcal{E}=E_{\bar{\mathcal{A}}}:=\{(j,\ell)\in E:\beta_j^*\beta_\ell^*=0\}$ (restriction to edges in the graph corresponding to pairs consisting of at least one zero coefficient). In the sequel, we denote by `Acc.E`, `Acc.E.A` and `Acc.E.Abar` the accuracy associated with $\mathcal{E}=E$, $\mathcal{E}=E_{\mathcal{A}}$ and $\mathcal{E}=E_{\bar{\mathcal{A}}}$ respectively.

Estimation and Prediction Performance. We also evaluate estimation consistency by computing the mean squared error (`MSE`) of the various estimators $\widehat{\boldsymbol{\beta}}$ based on empirical versions of $\mathbb{E}(\|\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}}\|^2)$. We finally evaluate Prediction accuracy on an external (test) sample. More precisely, after estimating $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$ on a dataset of size n, we generate an independent test sample of N observations $(\mathbf{z}_i^{(0)},Y_i^{(0)})$ (keep in mind that N is the largest sample size considered in a given scenario). Then, for every observation i of the test sample, we compute the predicted label $\widehat{Y}_i^{(0)}=\mathbb{I}(\text{logit}^{-1}(\mathbf{z}_i^{(0)T}\widehat{\boldsymbol{\beta}})>0.5)$, and the prediction accuracy is given by `Acc.Pred`$=(1/n)\sum_{i=1}^n\mathbb{I}(Y_i^{(0)}=\widehat{Y}_i^{(0)})$.

4.5 Illustration of our asymptotic results based on Zou's example

Sampling Covariates. Our first example is adapted from Zou (2006) with $p=4$ and $p_0=3$. Introduce $\rho_1=-0.39$ and $\rho_2=0.23$, $\mathbf{C}_{11}=1-\rho_1\mathbf{I}_3+\rho_1\mathbf{J}_3$ where $\mathbf{J}_3$ is the 3×3 matrix of 1's, and $C_{12}=\rho_2\mathbf{1}_3$. The covariance matrix $\mathbf{C}$ of the covariates is then defined as

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_{11} & C_{12}\\ C_{12}^T & 1\end{bmatrix}.$$

Zou (2006) used this example to illustrate situations where the Adaptive Lasso dramatically outperforms the crude Lasso (especially in terms of support recovery).

Setting $\boldsymbol{\beta}^*$. We set $\boldsymbol{\beta}^* \in \{\log(2.5), \log(2.5), \log(2.5), 0\}$.

Network generation. We consider a Generalized Fused-Lasso penalty that is based on a chain graph so that we have $\texttt{Spe}_G = \texttt{Sens}_G = 2/3$.

Setting p/n. To illustrate our asymptotic results we evaluate the methods for increasing sample sizes, namely $n = 30, 60, 120, 300, 600, 1200$ and 2000.

Results. Results are presented in Figure 2. First these results illustrate that Adaptive Generalized Fused-Lasso estimates are consistent (their MSE converge to 0 as n grows), and this consistency is slightly faster for Relaxed Adaptive Generalized Fused-Lasso estimates. They also tend to share the same support as $\boldsymbol{\beta}^*$ (Acc.A), and their successive differences also tend to share the same support as successive differences of $\boldsymbol{\beta}^*$ (Acc.E.A and Acc.E.Abar), when n grows. On the contrary non-Adaptive Generalized Fused-Lasso estimates do not enjoy these asymptotic oracle properties: they do not share the same support as $\boldsymbol{\beta}^*$ and are much slower to reach consistency (especially for the non-relaxed versions). Note that as expected the Relaxed Adaptive Lasso never shrinks successive differences of non-zero estimates to 0 so that this method is not efficient for the support recovery of successive differences (Acc.E.A is "artificially" high for low SNR values because estimates are all null, and so is their difference). More importantly, both consistency and sparsistency are slower for the Relaxed Adaptive Lasso, compared to (Relaxed) Adaptive Generalized Fused-Lasso. As for prediction accuracy, all the methods perform similarly for $n \geq 300$. For $n < 300$ however, Relaxed Adaptive Lasso is outperformed by all the versions of the Fused-Lasso, especially the ones using adaptive weights. As mentioned in Section 4, results obtained with the Adaptive Group Lasso are almost the same as those obtained with the Adaptive Lasso. As for GE-Net, results are a little worse because of the absence of adaptive weights. This simple example confirms that Fused-like estimates outperform Lasso estimates in some situations, in terms of both support recovery and prediction accuracy.

4.6 Assessing the performance of Adaptive Generalized Fused-Lasso estimates

Setting $\boldsymbol{\beta}^*$. in order to explore difficult/easy configurations in terms of SNR, we make non-null elements of $\boldsymbol{\beta}^*$ take values in $\{\log(1.1), \log(2), \log(4), \log(8), \log(12)\}$.

Sampling covariates. We sample $\mathbf{x}_i$ such that $\mathbf{x}_i \sim \mathcal{N}_p(0, \mathbf{C})$, $\mathbf{C}_{\text{AR}(1)} = 1/16 \left(\rho^{|i-j|}\right)_{i=1,p}^{j=1,p}$, with $\rho = -0.39$ as proposed in Zou (2006).

Setting p/n and p_0/p. We fix $p = 24$. As our theoretical results are asymptotic in n, we consider cases where $n/p \in \{1, 5, 10, 50\}$ to explore the performance of the methods in asymptotic and non-asymptotic settings. Moreover, we explore different values of ratio

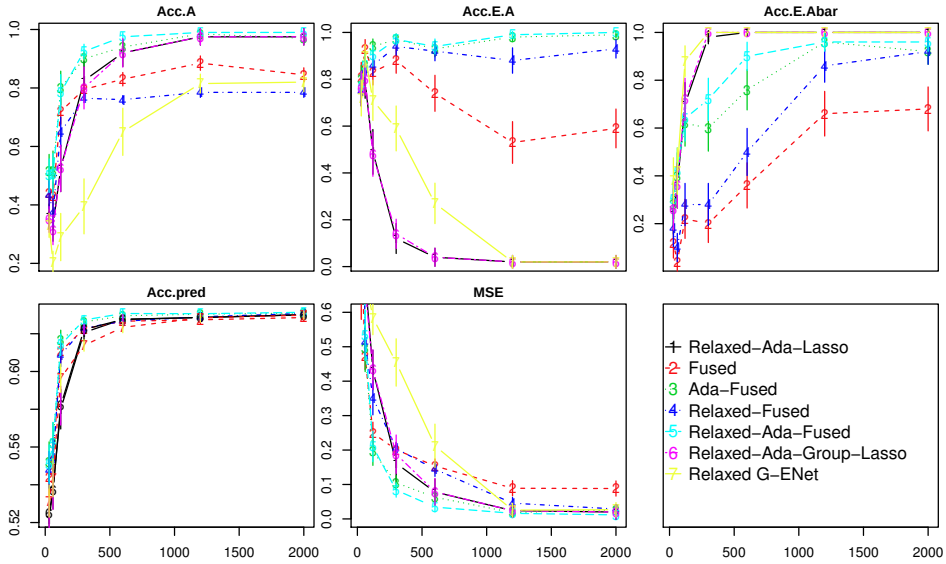


Figure 2: Comparison of the methods according to various criteria evaluated on Zou's example, for increasing sample sizes.

$p_0/p \in \{1/8, 1/4, 1/2\}$ as the number of *edges* connecting null and non-null coefficients depends on this ratio. Our results are summarized on Figures 4, 5, and 6.

Networks Generation for graph-based penalties. To study the influence of the graph in the Fused penalty, we generate networks of features with different topologies. To do so, we use a simplified version of the Stochastic Block Model (Airoldi et al., 2008), often called the affiliation graph model. In its simplest form this model considers two categories of nodes (null and non-null coefficients) whose connectivity is governed by parameter θ such that $\mathbb{P}\{j \sim k | j \in \mathcal{A}, k \in \bar{\mathcal{A}}\} = 1 - \theta$, $\mathbb{P}\{j \sim k | j \in \mathcal{A}, k \in \mathcal{A}\} = \theta$, and $\mathbb{P}\{j \sim k | j \in \bar{\mathcal{A}}, k \in \bar{\mathcal{A}}\} = \theta$, where $j \sim k$ for $j > k$ stands for $(j, k) \in E$. When θ increases there are fewer edges between null and non-null coefficients, the easiest configuration being when $\theta = 1$, the most difficult when $\theta = 0$. Parameter θ can also be interpreted in terms of Specificity and Sensitivity of the graph provided to the penalty since $\mathbb{E}(\texttt{Spe}_G(\theta)) = \mathbb{E}(\texttt{Sens}_G(\theta)) = \theta$. Examples of configurations considered here are given in Figure 3. Lastly, we mention that edges of the graph are fixed across replicates.

Result 1: ℓ_1-based Fused penalties show a cooperative effect on support recovery. When graph G is highly specific ($\theta = 1; 0.8$), ℓ_1-based fused penalties show better performance on support recovery compared with the *Relaxed Adaptive* Lasso (Figure 4). When applied to the difference of coefficients, the ℓ_1-norm helps in the identification of true zeros and true non-zeros based on informative edges. Of course, when the graph pro-

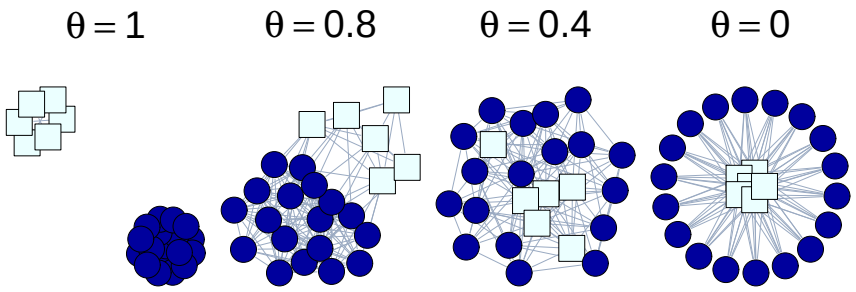


Figure 3: Networks generation for graph-based penalties with $p = 24$ and $p_0 = 6$. Two classes of nodes are present: rectangle white nodes correspond to null coefficients in $\boldsymbol{\beta}^*$ while colored circle nodes correspond to non-null coefficients. Parameter θ governs the intra/inter-group connectivity.

vides an erroneous information (edges between null and non-null coefficients only - $\theta = 0$), the *Relaxed Adaptive* Lasso should be preferred. Note that the "cooperative effect" of the ℓ_1-based fused penalty is higher when the proportion of non-null coefficients (p_0/p) is high, maybe because the proportion of informative edges increases with p_0/p.

Result 2: The "raw" Fused-Lasso is not robust to mis-specification of the graph in the penalty. The "raw" Fused-Lasso penalty (*i.e.* non-adaptive and without relaxation) shows poor performance when the graph is mis-specified. The Accuracy on support recovery is worst than the *Relaxed Adaptive* Lasso as soon as $\theta < 1$ (Fig. 4), and estimated parameters show an important bias even when $n/p = 50$ with a high SNR (Figure 6). Moreover, accuracies on pairs of coefficients (Figure 4) reflect the absence of selection, as the majority of coefficients are not shrunk to zero when using the Fused penalty only (Figure 6).

Result 3: Adaptive weights ? relaxation ? or both ? A general result is that using adaptive weights and/or relaxation is always preferable than the "raw" Fused-Lasso: *i*) Accuracies are higher (Acc.A, Acc.E.A, Acc.E.Abar), and *ii*) the modified versions of the Fused-Lasso are more robust to mis-specification of the graph. Conclusions are similar concerning the comparison with the *Relaxed Adaptive* Lasso that should be preferred when the specificity of the penalty is low. Then considering the *Adaptive* version of the Fused-Lasso (without relaxation), increases the accuracy support recovery that is higher than the Fused-Lasso, and increases also the accuracy on differences involving null coefficients (Acc.E.Abar). However, when focusing on non-null coefficients, some bias remains (even

	Acc.A		Acc.Pred	
	non-relaxed	relaxed	non-relaxed	relaxed
	$n/p = 1$			
non-adaptive	0.62	0.57	0.58	0.58
adaptive	0.54	0.55	0.58	0.58
	$n/p = 5$			
non-adaptive	0.67	0.73	0.61	0.62
adaptive	0.71	0.69	0.61	0.61
	$n/p = 10$			
non-adaptive	0.71	0.78	0.62	0.63
adaptive	0.78	0.76	0.63	0.63
	$n/p = 50$			
non-adaptive	0.79	0.88	0.65	0.65
adaptive	0.89	0.88	0.65	0.65

Table 3: Average Accuracies on support recovery (Acc.A) and on Prediction (Acc.Pred) according to the use of adaptive and/or relaxed Generalized Fused-Lasso estimates. Averages are computed across different configurations and SNRs.

if reduced, Fig. 6), and the relaxation strategy appears to be the most effective regarding bias reduction.

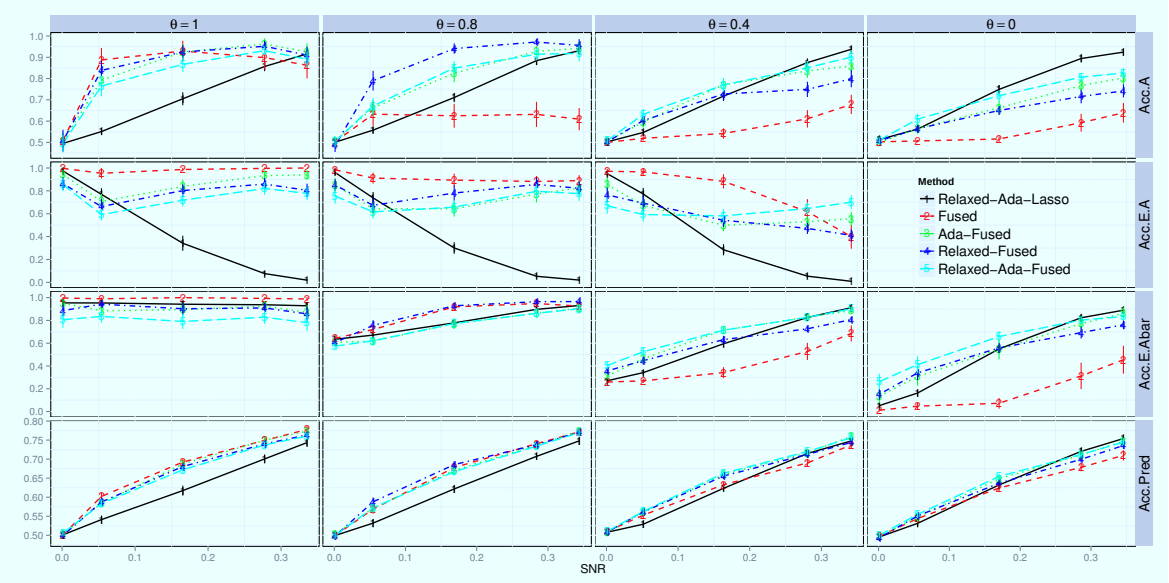
Lastly, we determine if considering both relaxation and adaptive weights can improve the performance of the Generalized Fused-Lasso. Table 3 provides a global view of variations of the Accuracies for support recovery according to n/p which allows us to draw a global conclusion (across specificities and SNRs):

i) when n/p is low (~ 1), neither the adaptive nor the relaxed Fused-Lasso outperform the "raw" Fused-Lasso, which suffices maybe because there is not enough available information to improve the selection procedure.

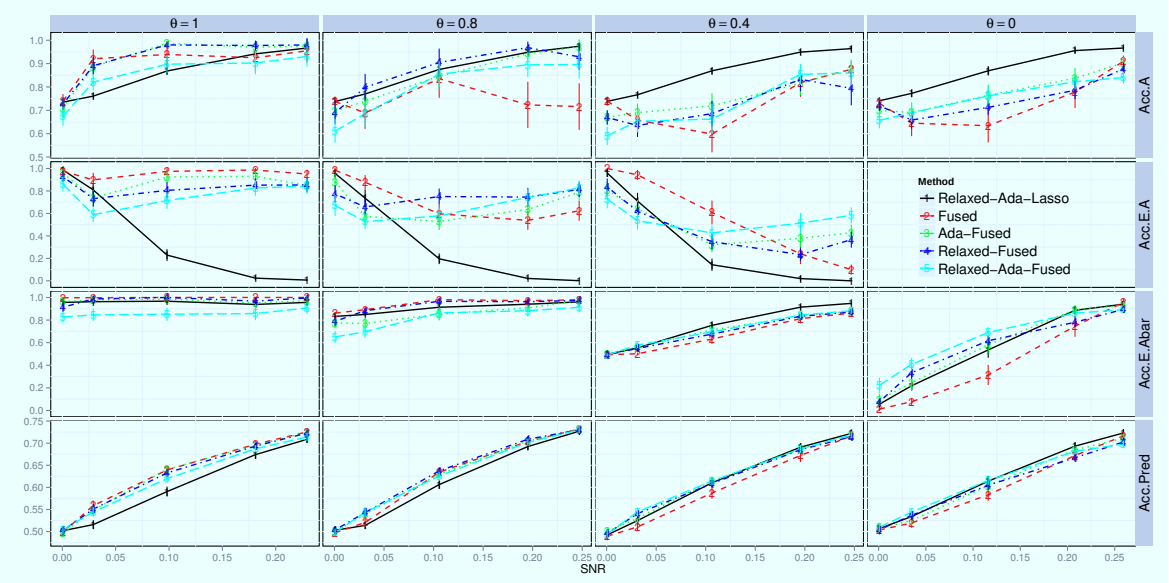
ii) when n/p is moderate ($\sim 5, 10$),using separate strategies is equivalent (relaxation without weights or weights without relaxation)

iii) when n/p is high (~ 50), the adaptive version is marginally more performant as more information is available for an accurate estimation of weights, and using relaxation does not change the performance.

iv) whatever n/p, all ℓ_1-based Fused penalties are more accurate than the Lasso for prediction (even when the graph provided is not specific, Fig. 5b), and there is no difference among them.

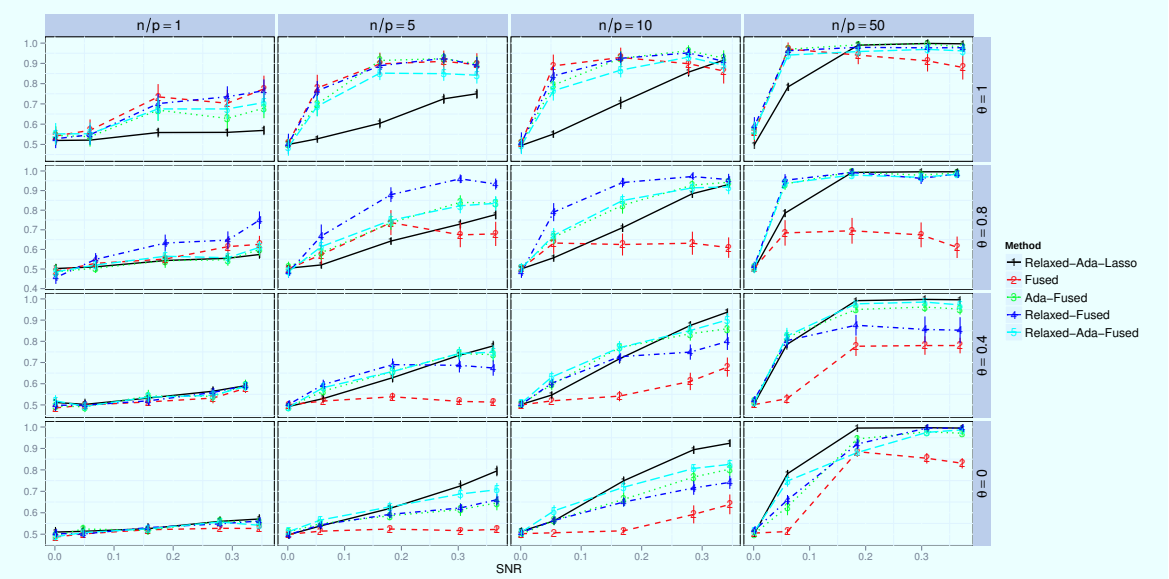


(a) Accuracies (support recovery and prediction) for ℓ_1-based Fused penalties ($n/p = 10$, $p_0 = p/2$).

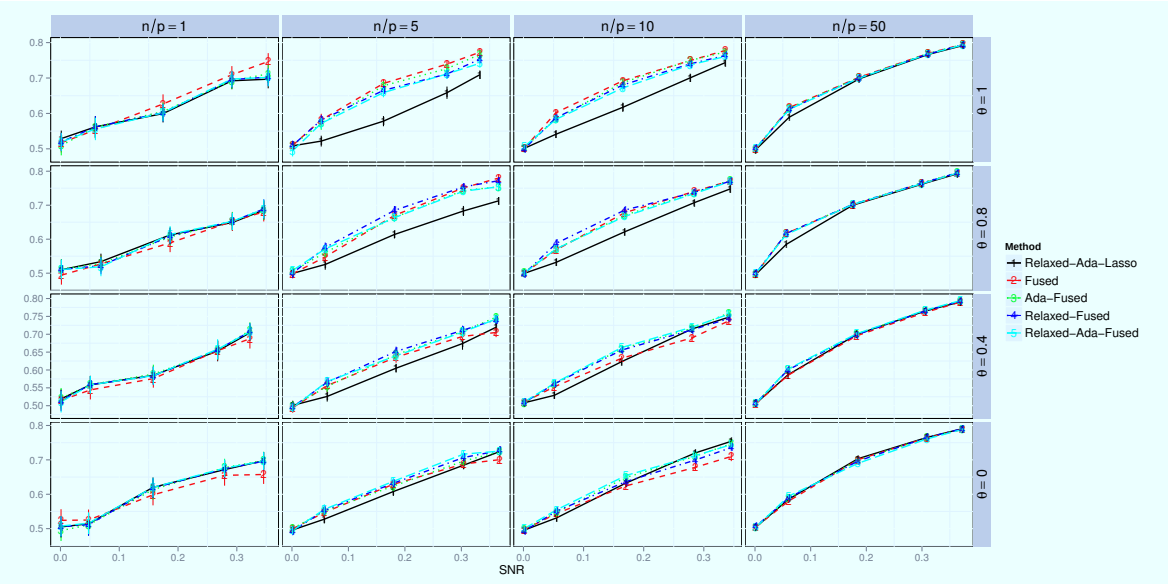


(b) Accuracies (support recovery and prediction) for ℓ_1-based Fused penalties ($n/p = 10$, $p_0 = p/4$).

Figure 4: Average accuracies for the detection of non-null coefficients (support recovery, Acc.A), and for the detection of null differences of coefficients without null coefficients (Acc.E.A) and with null-coefficients (Acc.E.Abar). The last row corresponds to the prediction accuracy (Acc.pred). Columns correspond to configurations ($\theta = 1, 0.8, 0.4, 0$). Results correspond to $p_0 = p/2$ (4a) and $p_0 = p/4$ (4b).

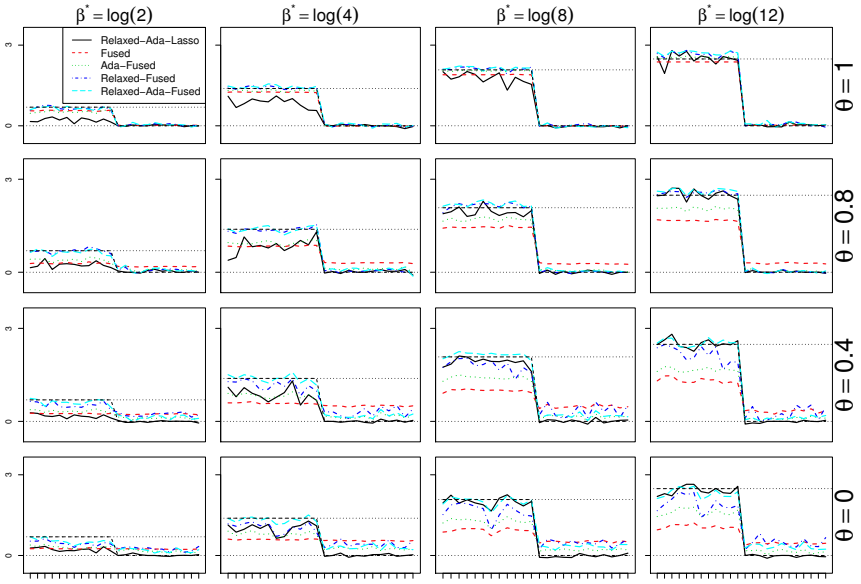


(a) Accuracies for support recovery (Acc.A) for Fused penalties with $p = 24$, and $n/p \in \{1, 5, 10, 50\}$, $p_0 = p/2$.

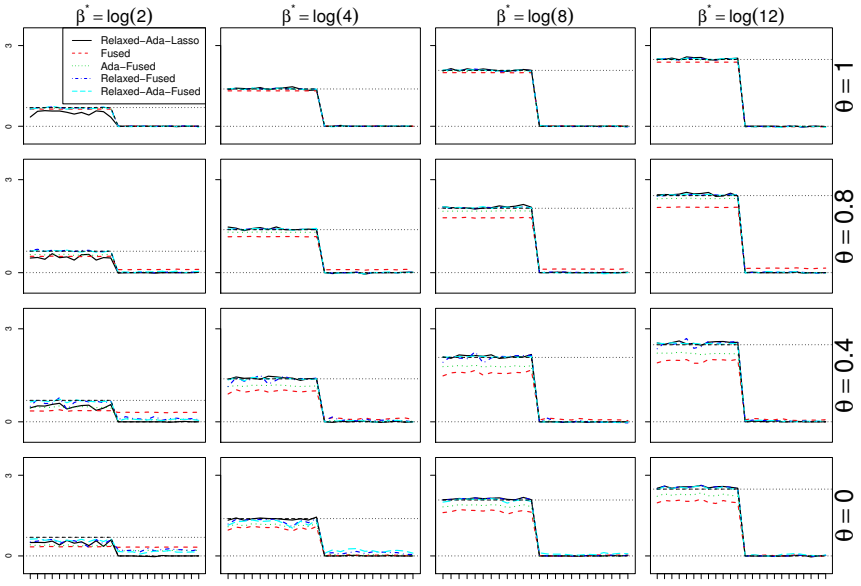


(b) Prediction accuracies for Fused penalties with $p = 24$, and $n/p \in \{1, 5, 10, 50\}$, $p_0 = p/2$.

Figure 5: Influence of the n/p ratio on average accuracies for the detection of non-null coefficients (support recovery, Acc.A, 5a), and on prediction accuracies, 5b). Rows correspond to configurations ($\theta = 1, 0.8, 0.4, 0$). Results correspond to $p_0 = p/2$.



(a) Estimated coefficients $\widehat{\beta}_j$ with $n/p = 10$.



(b) Estimated coefficients $\widehat{\beta}_j$ with $n/p = 50$.

Figure 6: Estimated coefficients with respect to different configurations. The x-axis corresponds to $p = 24$ positions of coefficients $\widehat{\boldsymbol{\beta}}$. Rows correspond to configurations $(\theta = 1, 0.8, 0.4, 0)$. Results correspond to $p_0 = p/2$.

4.7 Simulation Study in the context of Joint Modeling.

In this last simulation study we illustrate the interest of using Generalized Fused-Lasso estimates in the context of joint modeling, as described in Section 2.3. We start by setting the number of strata to $C = 4$ and we set $n_c = 200$ so that $n = \sum_c n_c = 800$. Those parameters remain fixed in the study.

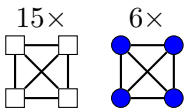
Setting β^*. Then we choose $p = 20$ and $p_0 = 6$ for each *stratum*, and non-null values of $\boldsymbol{\beta}_c^*$ take values in $\{\log(1.1), \log(1.5), \log(2), \log(2.5), \log(3), \log(5), \log(7), \log(9)\}$.

Sampling covariates. Covariates are generated for each stratum using a centered Gaussian distribution with covariance matrix $\mathbf{C}$ supposed to be the same for each stratum. Covariance matrix $\mathbf{C}$ is set to $\mathbf{C}_{\mathrm{AR}(1)}$ (as in the previous Section) with $\rho = 0.5$.

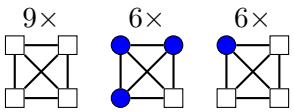
Varying Homogeneity between strata. The graph provided in the penalty is made of $p+1$ cliques of size C, each clique being used to connect coefficients $\{\beta_{j1}^*, \ldots, \beta_{jC}^*\}$, for $j = 0, \ldots, p$. Then we consider 3 different configurations that differ in the repartition of null and non-null coefficients within cliques, as illustrated in Figure 7. The first configuration corresponds to the case where nodes in the graph either correspond to 4 null or non-null theoretical coefficients (as in Fig. 7a). It is the most favorable configuration for Fused-like estimates ($\texttt{Spe}_G = 0.07$, $\texttt{Sens}_G = 1.00$). In the two other configurations true vectors $\boldsymbol{\beta}_c^*$ *do* vary across strata. The second design considers three types of cliques (Fig. 7b, $\texttt{Spe}_G = 0.05$ and $\texttt{Sens}_G = 0.97$), one being made of null coefficients only, and the other two mixing null and non-null in the same proportion (3/4). The third configuration is the most difficult one with cliques connecting half null and non-null coefficients (Fig. 7c, $\texttt{Spe}_G = 0.04$ and $\texttt{Sens}_G = 0.96$).

Competing methods. For the joint modeling problem, the Relaxed Adaptive Lasso (which corresponds to criterion (4) with $\lambda_n^{(2)} = 0$) is very similar to computing *Independent* Relaxed Adaptive Lasso estimates on each stratum. This latter option is more flexible because the sparsity parameter has not to be equal for each stratum. We therefore include it in our analysis instead of the Relaxed Adaptive Lasso: it is referred to as *Independent* Relaxed Adaptive Lassos (and denoted by Relaxed-Ada-Lasso-Indep on our figures). We also consider another version of the Lasso which consists in solving one Relaxed Adaptive Lasso on the whole data set (obtained by putting all the strata together), after selecting a reference stratum, and adding interaction terms between the remaining strata and the covariates. This method is described in more details in Section 7.5 of the Appendix. It is referred to as *Interaction* Relaxed Adaptive Lasso hereafter (and denoted by Relaxed-Ada-Lasso-Inter on our figures).

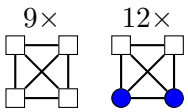
Results. Figures 8 presents the results. To make this figure more legible, only Adaptive versions of the Fused-Lasso estimates are presented: non-Adaptive versions achieved slightly worse performances in this simulation study. Our empirical findings confirm that configuration 1 is easier than configuration 2 which is itself easier than configuration 3, for



(a) Configuration 1: $\texttt{Spe}_G = 0.07$ and $\texttt{Sens}_G = 1.00$



(b) Configuration 2: $\texttt{Spe}_G = 0.05$ and $\texttt{Sens}_G = 0.97$



(c) Configuration 3: $\texttt{Spe}_G = 0.04$ and $\texttt{Sens}_G = 0.96$

Figure 7: Graphical representation of the three configurations considered in the simulation study for the joint modeling problem. Keep in mind that the graph on which the Fused penalty is based for this particular problem is composed by $p+1 = 21$ cliques of size $C = 4$. The figure presents every possible type of cliques for each configuration. Nodes correspond to coefficients, the theoretical value of which being either zero (white rectangles) or $\beta^* > 0$ (blue circles).

Fused-type estimates. For instance, regarding Fused-type estimates only, best [resp. worst] support recovery (Acc.A) and prediction (Acc.pred) accuracies are obtained for configuration 1 [resp. configuration 3]. As expected, *Independent* Relaxed Adaptive Lassos are are not sensitive to the level of heterogeneity across strata. But, for easy to moderately difficult configurations (configurations 1 and 2 here), Fused-type estimates (especially the Relaxed Adaptive Generalized Fused-Lasso ones) clearly outperform *Independent* Relaxed Adaptive Lasso estimates, in terms of overall support recovery, prediction accuracy and, of course, detection of heterogeneity across the strata (Acc.E.A and Acc.E.Abar). Under the most difficult configuration, there is no gain in using Fused-type estimates in terms of overall support recovery (for high SNR values, we even observe that *Independent* Relaxed Adaptive Lassos could achieve slightly better results). However, Relaxed Adaptive Generalized Fused-Lassos are at least comparable to *Independent* Relaxed Adaptive Lassos in terms of overall support recovery and still clearly better in terms of detection of heterogeneity across strata. Interestingly, they are also slightly better in terms of prediction accuracy (this result is consistent with what was also observed in Section 4.6 even in situations where the graph was poorly suited). Another interesting result is that the method relying on interaction terms is always outperformed by Fused-type estimates. Lastly, and overall, Relaxed Adaptive Generalized Fused-Lasso is slightly better [resp. worse] than its non-relaxed counterpart for support recovery, measured by Acc.A [resp. detection of heterogeneity across strata, measured by Acc.E.A].

5. Joint modeling to analyze road-safety data

Road safety is a major (political) concern in the West. Driving under the influence of alcohol (DUI) is an established risk factor of car accidents. Interestingly, several studies also suggest that DUI increases the risk of dying in an accident. But this result remains controversial: biological evidence supporting this assumption is still lacking and the observed effect of DUI on the risk of dying in an accident could be due to confounding variables

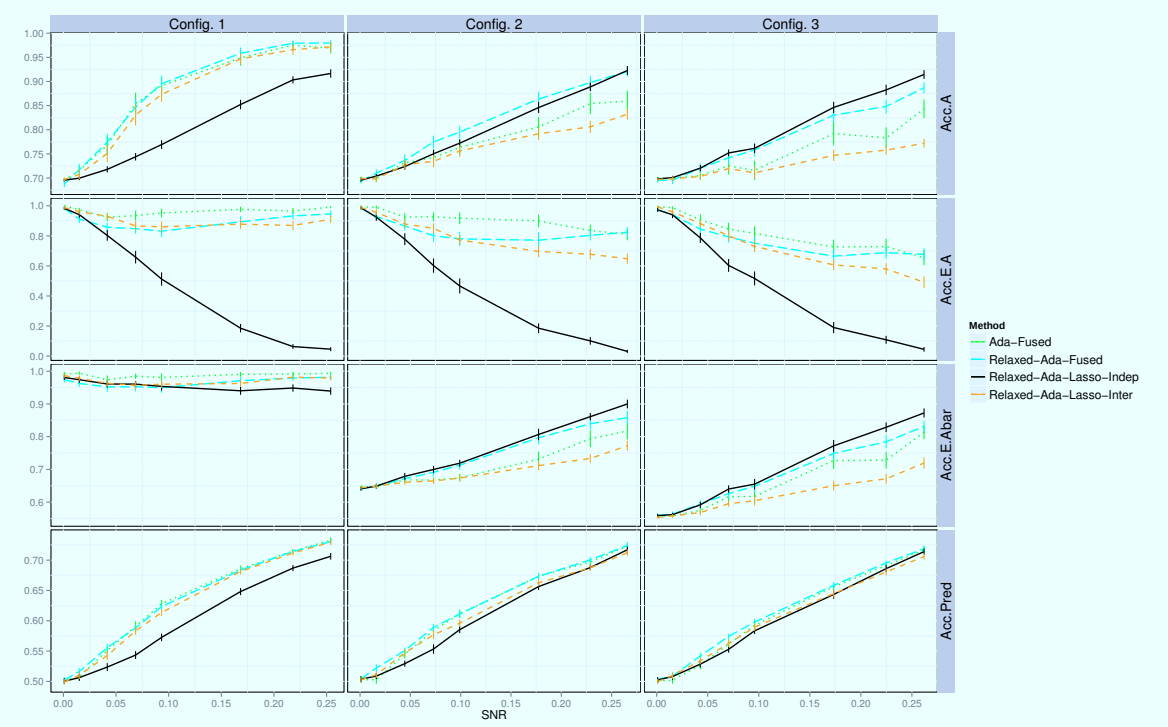


Figure 8: Comparisons of the methods in the joint modeling problem with dependent features ($\rho = 0.5$), according to various criteria. Each column corresponds to a given configuration, which is itself related to homogeneity between strata: from the most homogeneous (Configuration 1) to the most heterogeneous (Configuration 3).

only. In the following we propose a joint modeling approach to handle the effects of both DUI and confounding variables on the risk of dying in an accident.

Our dataset consists of $n = 21,064$ drivers involved in single-car personal injury crashes. The data were obtained from the systematic reports on road traffic injury crashes made by the police between 2006 and 2009 in France (metropolitan). Current data show 33 covariates including the characteristics of the crash: type of road, urban/rural location, lighting conditions (night with or without street lighting, dawn, daylight), and meteorological conditions (rain, wind, fog, fine weather) as well as the year of the crash. Data also include characteristics of the drivers such as gender, age, seat belt use and vital status (killed, injured or uninjured). Data on the crash-involved vehicles were also available, with the location of the main impact and the vehicle first registration year for instance. In the following we focus on the vital status of the driver only since every vehicle has one driver but a variable number of passengers. Lastly, we mention that crashes in which no driver was injured nor killed were excluded from the study, as well as crashes in which the only individuals to be injured or killed were passengers.

We define 4 strata based on gender and DUI: strata 1-2 for Males and Females not driving under the influence when the accident occured ($10,031$ and $6,385$ individuals respectively), and 3-4 for Males and Females driving under the influence when the accident occurred ($4,093$ and 555 individuals respectively). We use joint modeling to couple the estimation of the four logistic models relating the probability of dying in a car accident to various risk factors (one for each stratum). Most risk factors are indeed expected to share the same effect on the probability of dying across the four strata. In particular, we pay a particular attention to intercept parameters: they should be homogeneous across strata if neither gender nor alcohol directly modified the risk of dying in a car accident (gender and/or alcohol could have an indirect effect if other coefficients were observed to vary across strata). We compare different penalization strategies (Relaxed Adaptive Fused-Lasso, *Independent* Relaxed Adaptive Lassos, and *Interactions* Relaxed Adaptive Lasso) and we also present unpenalized estimates derived from standard logistic regression models independently built on each stratum. Each method returns a $\mathbb{R}^{34\times 4}$ matrix, where the four columns correspond to the four vectors of parameters $\hat{\boldsymbol{\beta}}_c \in \mathbb{R}^{34}$, including the intercept term, for strata $c = 1, \ldots, 4$ (Fig. 9).

Overall, there is a high concordance between supports of vectors $\hat{\boldsymbol{\beta}}_c$, $c = 1, \ldots, 4$, returned by penalized methods, especially Relaxed Adaptive Generalized Fused-Lasso and *Independent* Relaxed Adaptive Lassos. For instance, penalized methods all agree on the absence of significant influence of most recorded variables on the risk of dying in an accident, given an accident occurred (the driver's professional activity, the reason the journey -professional or not- the period of the year -Q1, Q2 and Q3- i.e. the first, second and third thirds of the year, the period of the day, the vehicle age, the year of crash, atmospheric

conditions, the adherence condition, the flatness or the curvature of the road). On the other hand, the risk appears to be higher for older drivers and for drivers under drugs, as well as if the accident occurred on a shoulder, if the car hit a wall or a tree and if the main impact was on the left side of the car. Factors that reduce the risk of dying are the use of a seat-belt, non-bidirectional roads (i.e., roads with a median strip) and city roads (where speed is generally lower). Interestingly the Relaxed Adaptive Fused-Lasso and *Independent* Relaxed Adaptive Lassos return slightly different estimates for the influence of city roads. This can be explained by aliasing between covariate "City Roads" (that equals 1 if the road section is managed by a city, and not by a county, a region nor the country), and covariate "City" that equals one if the crash occurred in a city (but not necessarily on a road section managed by a city). These two covariates are correlated since covariate "City'" is most often 1 when "City Roads" is 1. It would make sense to exclude one of them but we keep both for illustration purposes. According to the Relaxed Adaptive Fused, both covariates have homogeneous effects across strata, which suggests no interaction between these covariates and gender and DUI. This is also suggested by the *Interactions* Relaxed Adaptive Lasso. As for independent Lasso estimates, they suggest differences across strata, which are hard to interpret in this case. There are a few other differences between the three penalized methods, but when Relaxed Adaptive Fused-Lasso and *Interactions* Relaxed Adaptive Lasso estimates disagree, the Relaxed Adaptive Fused-Lasso most often agrees with *Independent* Relaxed Adaptive Lassos. This is consistent with our observations from the simulation study where these latter two methods performed the best. Let us finally mention that the structure of the matrix obtained with the Adaptive Fused Lasso which was the best for the detection of heterogeneity across strata on our simulation study, was 100% consistent with the one obtained with the Relaxed Adaptive Fused-Lasso.

Finally, the most important result is that Relaxed Adaptive Fused-Lasso indicates that intercepts *do* vary across strata, suggesting an effect of both gender and DUI on the risk of dying in a single-crash accident. More precisely, sober females are at a higher risk than sober males, and, to a lesser extent, females under the influence are at a higher risk than males under the influence. Moreover, irrespective on gender, drivers under the influence are at a higher risk than sober drivers. However this result should be tempered by potential confounding due to speed, which was not available here. For instance, drivers under the influence are likely to drive faster than sober drivers. The effects of speed may be partly captured by other covariates, but not entirely. Consequently residual "speed effects" could be responsible for detected heterogeneities between intercepts.

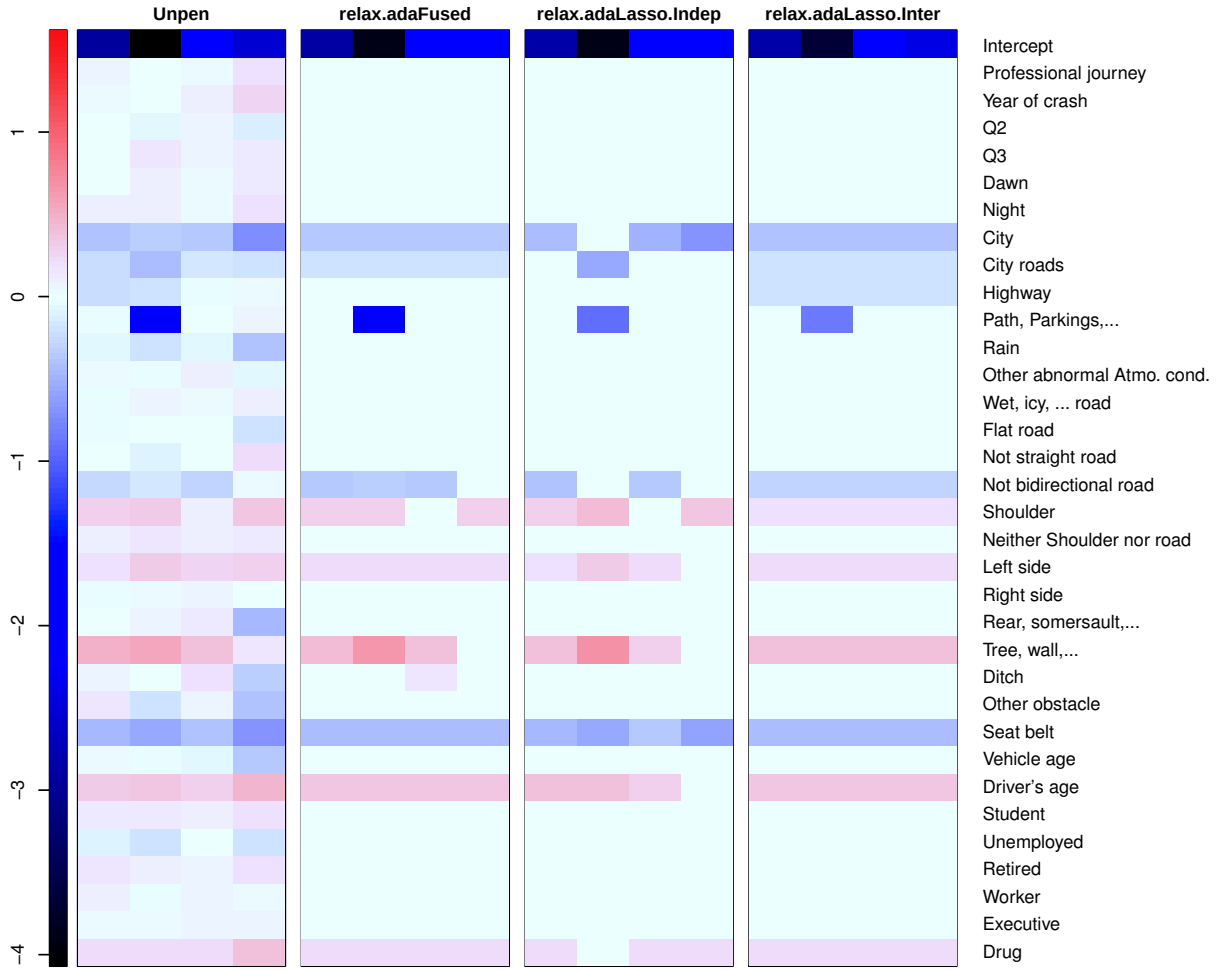


Figure 9: Estimates of the logistic regression models for studying the risk a dying in a car accident with strata defined by combining gender and alcohol consumption.

6. Discussion

In this work we established the asymptotic oracle properties of the Adaptive Generalized Fused-Lasso estimates in the case of linear and logistic regression models, for fixed p. To our knowledge these results are the first established for Fused-Lasso estimates in the setting of GLMs, and are also the first established for the Generalized Fused penalty based on a graph. Of course these results should be extended especially to the the high-dimensional case where p may grow with n (or even $p \geq n$). As mentioned in the Introduction, most published papers dealing with Fused-Lasso estimates in high-dimension focused on the chain-based Fused penalty in the Gaussian sequence model (*i.e.* the original Fused Lasso of Tibshirani (1996) that supposes an identity design matrix $\mathbf{X}$). Vaiter et al. (2011) recently established some results for a class of penalized least-squares based on a class of ℓ_1 penalties that encompasses Generalized Fused Lassos. However the extension of such results to GLMs with Adaptive weights would not be straightforward and thus would be an interesting lead. Above all, non-asymptotic oracle prediction inequalities still need to be established for Fused-like estimates: under the Linear Model for instance, and setting $\widehat{\boldsymbol{\beta}}$ the Fused-Lasso estimate for some appropriate values of the tuning parameters, it is easy to obtain inequalities of the form

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \leq \kappa \frac{p_0 \log(p)}{n} \quad \text{with high probability,}$$

for some positive constant κ, but the main question is about getting a similar inequality with p_0 replaced by the quantity s_0 introduced in Section 3. Rinaldo (2009) established such an inequality in the Gaussian sequence model, and only for a modification of the Fused-Lasso (that was called Adaptive Fused-Lasso even though it does not rely on adaptive weights). We believe that our simulation study suggests that such an inequality should hold for Adaptive Generalized Fused-Lasso estimates, even if it was conducted in the low-dimensional case. Indeed, we observed very good prediction performance for these estimates under the Logistic Model. In particular, when the graph is appropriately chosen, predictive performance of Fused-like estimates are much higher than those of the Lasso for moderate sample sizes.

From the modeling point of view, using a graph that provides a correct information in the penalty is of course crucial, and we demonstrated that the performance of Adaptive Generalized Fused-Lasso estimates is deeply related to the suitability of the chosen graph, especially for support recovery. This graph constitutes a formal description of some *prior* knowledge on the problem that is investigated. We may stress that this graph does not describe correlations among features but similarity between the effects of these features under some model. Correlated features may share similar coefficients but this is not always the case. For instance in epidemiology, smoking and alcohol consumption are generally highly correlated. They further may share similar effects under a logistic model when studying cardiovascular diseases so that it might make sense to draw an edge between the

corresponding components in a graph. However, if studying lung cancer, they would not share similar coefficients at all, and such an edge should not be drawn.

A particular situation where the graph is suggested by the design of the study itself is what was called Joint Modeling here, where data come from various strata. In the most general case, the graph consists of $p+1$ cliques, whose common size is the number of strata. When the main question of interest is the detection of heterogeneity across the strata, we believe that this type of graph is very natural. Indeed, it encourages coefficients to be homogeneous across the strata, so that detected heterogeneities come from the data. It has some connections with the statistical tests theory where tests are generally performed under the null hypothesis (absence of heterogeneity in this case), and data need to be far from this assumption in order to reject the null hypothesis. But even in the Joint Modeling context, other graphs may be considered, under some particular circumstances: for instance, stars may replace cliques if one stratum can serve as a reference.

7. Appendix

7.1 Application to the joint modeling of sparse regression models

Our objective here is to show how to rewrite the joint modeling problem as a particular case of the Generalized Fused-Lasso problem introduced in (4). With the same notations as in Section 2.3, the criterion considered in the joint modeling context can be written,

$$\sum_{c=1}^{C}\Big\{\sum_{i:\mathcal{C}_i=c}\mathcal{J}(Y_i,\mathbf{z}_i^T\boldsymbol{\beta}_c)+\lambda_n^{(1)}\sum_{j=1}^{p}w_j^{(1)}|\beta_{c,j}|\Big\}+\lambda_n^{(2)}\sum_{j=0}^{p+1}\sum_{c_1>c_2}w_{c_1,c_2,j}^{(2)}|\beta_{c_1,j}-\beta_{c_2,j}|, \qquad (5)$$

where $\mathcal{J}$ can typically be set to either $\mathcal{J}_{\text{sq}}$ or $\mathcal{J}_{\text{lo}}$ depending on whether linear or logistic models are considered (see (1) and (2) above). In the absence of the $\lambda_n^{(2)}$-penalty term, optimizing this criterion reduces to independently solve C Lasso problems, one for each stratum. The $\lambda_n^{(2)}$-penalty term encourages coefficients $\beta_{c_1,j}$ and $\beta_{c_2,j}$ (that is, the coefficient of the j-th covariate in the c_1-th and c_2-th stratum respectively) to be equal or at least close to each other: in this sense, it couples the Lasso estimates that would be obtained without this $\lambda_n^{(2)}$-penalty. In (5) above, intercept terms are not directly penalized (the sum involved in the $\lambda_n^{(1)}$-penalty starts at $j=1$) but are encouraged to be equal (the sum involved in the $\lambda_n^{(1)}$-penalty starts at $j=0$). Of course, other options for the intercept terms can be considered.

We now show how to rewrite criterion (5) as a version of (4). Towards this aim, set $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_C^T)^T\in\mathbb{R}^{C(p+1)}$ and denote by $\mathcal{Z}=\text{diag}(\mathbf{Z}_1,\ldots,\mathbf{Z}_C)$, the $[n\times C(p+1)]$ block-diagonal matrix, whose blocks are of size $[n_c\times(p+1)]$ and defined as $\mathbf{Z}_c=[\mathbf{1}\ \mathbf{X}_c]$. Here $\mathbf{X}_c$ stands for the $[n_c\times p]$ sub-matrix of $\mathbf{X}$ with rows corresponding to observations falling in stratum c (by assumption, the vector of response variables is given by $\mathbf{Y}=(\mathbf{Y}_1^T,\ldots,\mathbf{Y}_C^T)^T$ with $\mathbf{Y}_c:=\{Y_i,\mathcal{C}_i=c\}$ denoting the vector of length n_c containing the response variables

corresponding to observations of the c-th stratum). Letting, for any two integers n_1, n_2, $n_1 \% n_2$ be the rest of the division of n_1 by n_2, criterion (5) then writes

$$\sum_{i=1}^{n} \mathcal{J}(Y_i, \mathcal{Z}_i^T \boldsymbol{\beta}) + \lambda_n^{(1)} \sum_{\substack{j=0 \\ j\%(p+1)\neq 0}}^{C(p+1)} w_j^{(1)} |\beta_j| + \lambda_n^{(2)} \sum_{(j,\ell)\in E} w_{j\ell}^{(2)} |\beta_j - \beta_\ell|, \tag{6}$$

where the edge set E along with the node set $V = \{1, \ldots, C(p+1)\}$ define a graph $G = (V, E)$ which consists of $p+1$ cliques of size C (see Figure 1). Edges (j, ℓ) present in graph G are those for which $j > \ell$ and $\ell\%(p+1) = j\%(p+1)$. Criterion (6) is exactly criterion (4) with $\mathcal{Z}$ instead of $\mathbf{Z}$, and some terms, corresponding to intercept parameters, that are not penalized in the $\lambda_n^{(1)}$-penalty.

In some situations, one stratum can be regarded as the reference: for instance, when strata correspond to various treatments, the control treatment can serve as the reference. Without loss of generality, we can assume that the reference stratum is the first one, in which case an adapted version of penalty (5) is

$$\sum_{c=1}^{C} \left\{ \lambda_n^{(1)} \sum_{j=2}^{p+1} |\beta_{c,j}| \right\} + \lambda_n^{(2)} \sum_{j=1}^{p+1} \sum_{c=2}^{C} |\beta_{c,j} - \beta_{1,j}|.$$

The graph induced by this penalty consist of $p+1$ star graphs (one for each covariate plus one for the intercept) with the reference at its center: more precisely, in the j-th star graph, the only present connections are between coefficient $\beta_{1,j}$ and the $C-1$ remaining coefficients $\beta_{c,j}$, for $c \neq 1$. An illustration is presented in the right panel of Figure 1, in the case $p = 2$ and $C = 4$.

7.2 Proof of Theorem 1

This proof is an adaptation to the logistic case of the proof given by Tibshirani et al. (2005). The major difference concerns the treatment of the loss function. Let us define $\mathcal{V}_n(\mathbf{u}) = Q(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}) - Q(\boldsymbol{\beta}^*)$ with $\mathbf{u} = (u_0, \ldots, u_p)^T$, and Q defined as in criterion (4) with $J = J_{\text{lo}}$. Obviously $\mathcal{V}_n(\mathbf{u})$ is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Similarly to Tibshirani et al. (2005), we obtain:

$$\begin{aligned}\mathcal{V}_n(\mathbf{u}) = J_{\text{lo}}\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right) - J_{\text{lo}}(\boldsymbol{\beta}^*) + \lambda_n^{(1)} \sum_{j=1}^{p} \left\{ \left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_j^*| \right\} \\ + \lambda_n^{(2)} \sum_{(j,\ell)\in E} \left\{ \left|\beta_j^* - \beta_\ell^* + \frac{(u_j - u_\ell)}{\sqrt{n}}\right| - |\beta_j^* - \beta_\ell^*| \right\}.\end{aligned}$$

For any fixed $\mathbf{u}$, the last two terms of the right-hand side converge to the last two terms in expression (5) of $\mathcal{V}(\mathbf{u})$ as n goes to ∞. As for the first two terms, a Taylor expansion

yields

$$J_{\text{lo}}\left(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}\right) - J_{\text{lo}}(\boldsymbol{\beta}^*) = \nabla J_{\text{lo}}(\boldsymbol{\beta}^*)^T \frac{\mathbf{u}}{\sqrt{n}} + \frac{1}{2}\mathbf{u}^T \frac{\mathcal{I}(\boldsymbol{\beta}^*)}{n}\mathbf{u} + o_{\mathbb{P}}(1/n).$$

But, under **AL1**, we have

$$\frac{1}{2}\mathbf{u}^T \frac{\mathcal{I}(\boldsymbol{\beta}^*)}{n}\mathbf{u} \to_d \frac{1}{2}\mathbf{u}^T \mathbf{C}\mathbf{u}.$$

Moreover, **AL1** implies that the minimum eigenvalue of $\mathcal{I}(\boldsymbol{\beta}^*)$ goes to ∞ and, under **AL2**, it is well known (Gourieroux and Monfort (1981)) that

$$\frac{\nabla J_{\text{lo}}(\boldsymbol{\beta}^*)}{\sqrt{n}} \to_d \mathbf{W}, \tag{7}$$

where $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}_{p+1}, \mathbf{C})$ distribution. By Slutsky's theorem, we therefore have $\mathcal{V}_n(\mathbf{u}) \to_d \mathcal{V}(\mathbf{u})$. Since $\mathcal{V}_n$ is convex, the epi-convergence results of Geyer (1994) can finally be used to complete the proof of Theorem 1.

7.3 Proof of Proposition 2

For the sake of brevity, the proof is only given in the logistic case; the Gaussian case follows from the exact same lines. If $\lambda_0^{(2)} = 0$, the proof is the same as in the "pure" Lasso case (see Zou (2006)). If $\lambda_0^{(2)} \neq 0$, the result follows from an adaptation of Zou's proof. Below, we assume that $\lambda_0^{(1)} \neq 0$; the case where $\lambda_0^{(1)} = 0$ is slightly easier and omitted. First observe that $\mathbb{P}(\widetilde{\mathcal{A}}_n = \mathcal{A}) \leq \mathbb{P}(\sqrt{n}\hat{\beta}_j = 0 \ \forall j \notin \mathcal{A})$. Moreover, in virtue of Theorem 1, we have $\limsup_n \mathbb{P}(\sqrt{n}\hat{\beta}_j = 0 \ \forall j \notin \mathcal{A}) \leq \mathbb{P}(u_j^* = 0 \ \forall j \notin \mathcal{A})$. Therefore, we only need to show that $c = \mathbb{P}(u_j^* = 0 \ \forall j \notin \mathcal{A}) < 1$.
For any $j \in \{1, \ldots, p\}$, introduce $E_j^{=}(\boldsymbol{\beta}^*) = \{\ell : (\ell, j) \in E^{=}(\boldsymbol{\beta}^*) \text{ or } (j, \ell) \in E^{=}(\boldsymbol{\beta}^*)\}$ and $E_j^{\neq}(\boldsymbol{\beta}^*) = \{\ell : (\ell, j) \in E^{\neq}(\boldsymbol{\beta}^*) \text{ or } (j, \ell) \in E^{\neq}(\boldsymbol{\beta}^*)\}$, where the sets of pairs of vertices $E^{=}(\boldsymbol{\beta}^*)$ and $E^{\neq}(\boldsymbol{\beta}^*)$ were introduced in Section 4.4. Let $\mathbf{u}^* = \text{argmin}(\mathcal{V})$. Setting $\mathbf{W} = (W_0, \ldots, W_p)^T$ and $\mathbf{C}\mathbf{u}^* = ((\mathbf{C}\mathbf{u}^*)_0, \ldots, (\mathbf{C}\mathbf{u}^*)_p)^T$, we have, by the Karush-Kuhn-Tucker (KKT) optimality condition,

$$W_0 + (\mathbf{C}\mathbf{u}^*)_0 = 0, \tag{8}$$

and for all $j \in \mathcal{A}$,

$$W_j + (\mathbf{C}\mathbf{u}^*)_j + \lambda_0^{(1)} sign(\beta_j^*) + \lambda_0^{(2)}\Big\{ \sum_{\ell \in E_j^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_j^* - \beta_k^*) + \sum_{\ell \in E_j^{=}(\boldsymbol{\beta}^*)} (-1)^{\mathbb{I}(j<\ell)} t_{j\ell} \Big\} = 0,$$

and for all $j \notin \mathcal{A}$,

$$W_j + (\mathbf{C}\mathbf{u}^*)_j + \lambda_0^{(1)} r_j + \lambda_0^{(2)}\Big\{ \sum_{\ell \in E_j^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_j^* - \beta_k^*) + \sum_{\ell \in E_j^{=}(\boldsymbol{\beta}^*)} (-1)^{\mathbb{I}(j<\ell)} t_{j\ell} \Big\} = 0.$$

Above, $r_j = sign(u_j^*)$ if $u_j^* \neq 0$ and r_j is some real number in $[-1,1]$ otherwise. Similarly, $t_{j\ell} = sign(u_j^* - u_\ell^*)$ if $u_j^* \neq u_\ell^*$ and $t_{j\ell}$ is some real number in $[-1,1]$ otherwise. For any index $j \in \mathcal{A}$ there is some $s = s(j)$ such that $j \in \mathcal{A}_{s(j)}$, where $\mathcal{A}_s$ still denotes the set of vertices of the s-th connected component of $G_{\mathcal{B}}$ (see the paragraph before the statement of Theorem 3 for the definitions of these objects). Then summing up the KKT optimality conditions over the set $\mathcal{A}_{s(j)}$, we have

$$\sum_{k \in \mathcal{A}_{s(j)}} \Big\{ W_k + (\mathbf{C}\mathbf{u}^*)_k + \lambda_0^{(1)} sign(\beta_k^*) + \lambda_0^{(2)} \sum_{\ell \in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^* - \beta_\ell^*) \Big\} = 0. \tag{9}$$

Similarly, setting $\widetilde{\mathcal{B}} = \{(j,\ell) \in E \cap \bar{\mathcal{A}} \times \bar{\mathcal{A}}\}$ and denoting by $G_0 = (\bar{\mathcal{A}}, \widetilde{\mathcal{B}})$, the set $\bar{\mathcal{A}}$ can be decomposed as $\bar{\mathcal{A}} = \cup_{s=1}^{s_1} \bar{\mathcal{A}}_s$, where $1 \leq s_1 \leq p - p_0$ and $\bar{\mathcal{A}}_s$ is the subset of vertices constituting the s-th connected component of G_0. Then, for any $j \notin \mathcal{A}$, there exists some $s = s(j)$ such that $j \in \bar{\mathcal{A}}_{s(j)}$ and summing up the KKT optimality conditions over $\bar{\mathcal{A}}_{s(j)}$, we have

$$\sum_{k \in \bar{\mathcal{A}}_{s(j)}} \Big\{ W_k + (\mathbf{C}\mathbf{u}^*)_k + \lambda_0^{(1)} r_k + \lambda_0^{(2)} \sum_{\ell \in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^* - \beta_\ell^*) \Big\} = 0, \tag{10}$$

with $|s_k| \leq 1$. If $u_j^* = 0$ for all $j \notin \mathcal{A}$, equations (9) along with equation (8) form a system of $s_0 + 1$ equations with $p_0 + 1 \geq s_0 + 1$ variables, that can be written as

$$\mathbf{W}_{\mathcal{B}} + \mathbf{M}_1 \mathbf{u}^*_{\{0\} \cup \mathcal{A}} + \lambda_0^{(1)} \mathbf{r}_{\mathcal{B}} + \lambda_0^{(2)} \mathbf{t}_{\mathcal{B}} = \mathbf{0},$$

where $\mathbf{M}_1$ is the $(s_0+1) \times (p_0+1)$ matrix whose (s,j) element is $m_{s,j} = \sum_{k \in \mathcal{A}_{s-1}} C_{k,j}$ (with $C_{k,j}$ the (k,j) element of $\mathbf{C}$ and $\mathcal{A}_0 = \{0\}$), and $\mathbf{W}_{\mathcal{B}}, \mathbf{r}_{\mathcal{B}}$ and $\mathbf{t}_{\mathcal{B}}$ are vectors in $\mathbb{R}^{s_0+1}$ whose s-th elements are $\sum_{k \in \mathcal{A}_{s-1}} W_k$, $\sum_{k \in \mathcal{A}_{s-1}} sign(\beta_k^*)$ and $\sum_{k \in \mathcal{A}_{s-1}} \sum_{\ell \in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^* - \beta_\ell^*)$ respectively. Now, denoting by $\mathbf{M}_1^\dagger$ the pseudo-inverse of $\mathbf{M}_1$, there exists some vector $\boldsymbol{\omega} \in \mathbb{R}^{p_0+1}$ such that

$$\mathbf{u}^*_{\{0\} \cup \mathcal{A}} = (\mathbf{I}_{p_0+1} - \mathbf{M}_1^\dagger \mathbf{M}_1)\boldsymbol{\omega} - \mathbf{M}_1^\dagger (\mathbf{W}_{\mathcal{B}} + \lambda_0^{(1)} \mathbf{r}_{\mathcal{B}} + \lambda_0^{(2)} \mathbf{t}_{\mathcal{B}}). \tag{11}$$

On the other hand, if $u_j^* = 0$ for all $j \notin \mathcal{A}$, equations (10) form a system of s_1 equations that can be written, in view of (11),

$$\Big| \mathbf{W}_{\widetilde{\mathcal{B}}} + \lambda_0^{(2)} \mathbf{t}_{\widetilde{\mathcal{B}}} + \mathbf{M}_2 \big\{ (\mathbf{I}_{p_0} - \mathbf{M}_1^\dagger \mathbf{M}_1)\boldsymbol{\omega} - \mathbf{M}_1^\dagger (\mathbf{W}_{\mathcal{B}} + \lambda_0^{(1)} \mathbf{r}_{\mathcal{B}} + \lambda_0^{(2)} \mathbf{t}_{\mathcal{B}}) \big\} \Big| \leq \lambda_0^{(1)} \mathbf{r}_{\widetilde{\mathcal{B}}},$$

where $\mathbf{W}_{\widetilde{\mathcal{B}}}, \mathbf{r}_{\widetilde{\mathcal{B}}}$ and $\mathbf{t}_{\widetilde{\mathcal{B}}}$ are the vectors in $\mathbb{R}^{s_1}$ whose s-th elements are given by $\sum_{k \in \bar{\mathcal{A}}_s} W_k$, $|\bar{\mathcal{A}}_s|$ and $\sum_{k \in \bar{\mathcal{A}}_s} \sum_{\ell \in E_k^{\neq}(\boldsymbol{\beta}^*)} sign(\beta_k^* - \beta_\ell^*)$ respectively. We can now conclude by observing that

$$c \leq \mathbb{P}\Big(\Big| \mathbf{W}_{\widetilde{\mathcal{B}}} + \lambda_0^{(2)} \mathbf{t}_{\widetilde{\mathcal{B}}} + \mathbf{M}_2 \big\{ (\mathbf{I}_{p_0} - \mathbf{M}_1^\dagger \mathbf{M}_1)\boldsymbol{\omega} - \mathbf{M}_1^\dagger (\mathbf{W}_{\mathcal{B}} + \lambda_0^{(1)} \mathbf{r}_{\mathcal{B}} + \lambda_0^{(2)} \mathbf{t}_{\mathcal{B}}) \big\} \Big| \leq \lambda_0^{(1)} \mathbf{r}_{\widetilde{\mathcal{B}}} \Big) < 1.$$

7.4 Proof of Theorem 3

For the sake of brevity, the proof is solely detailed in the Gaussian case. The proof in the logistic case mostly follows from similar arguments along with others used to derive Theorem 1. We will briefly describe how to proceed when the Gaussian case differs from the logistic one.

The following proof is an adaptation of the proof given by Zou (2006). The main difference concerns the treatment of the penalty term. Let us define $\mathbf{V}_n(\mathbf{u}) = Q(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}) - Q(\boldsymbol{\beta}^*)$ with $\mathbf{u} = (u_0, \ldots, u_p)^T$ and Q defined as in (4) with $J = J_{\text{sq}}$. Note that $\mathbf{V}_n(\mathbf{u})$ is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ad} - \boldsymbol{\beta}^*)$. We have

$$\begin{aligned}\mathbf{V}_n(\mathbf{u}) = \mathbf{u}^T \left(\frac{1}{2n}\mathbf{Z}^T\mathbf{Z}\right)\mathbf{u} - \frac{\epsilon^T\mathbf{Z}}{\sqrt{n}}\mathbf{u} + \frac{\lambda_n^{(1)}}{\sqrt{n}}\sum_{j=1}^{p} w_j^{(1)}\sqrt{n}\left\{\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_j^*|\right\} + \\ \frac{\lambda_n^{(2)}}{\sqrt{n}}\sum_{(j,\ell)\in E} w_{j\ell}^{(2)}\sqrt{n}\left\{\left|\beta_j^* - \beta_\ell^* + \frac{(u_j - u_\ell)}{\sqrt{n}}\right| - |\beta_j^* - \beta_\ell^*|\right\} \\ =: \mathbf{u}^T\left(\frac{1}{n}\mathbf{Z}^T\mathbf{Z}\right)\mathbf{u} - 2\frac{\epsilon^T\mathbf{Z}}{\sqrt{n}}\mathbf{u} + \sum_{j=1}^{p} T_j^{(1)} + \sum_{(j,\ell)\in E} T_{j\ell}^{(2)}\end{aligned}$$

We have the two following behaviors :

$$T_j^{(1)} \to_p \begin{cases} 0 & \text{if } \beta_j^* \neq 0 \text{ or } (\beta_j^* = 0 \text{ and } u_j = 0) \\ \infty & \text{otherwise} \end{cases}$$

and

$$T_{j\ell}^{(2)} \to_p \begin{cases} 0 & \text{if } \beta_j^* \neq \beta_\ell^* \text{ or } (\beta_j^* = \beta_\ell^* \text{ and } u_j = u_\ell) \\ \infty & \text{otherwise} \end{cases}.$$

Denote by $\mathbf{C}_{\mathcal{A}}$ the $(p_0+1)\times(p_0+1)$ sub-matrix of $\mathbf{C}$ constituted of rows and columns associated with indexes in $\{0\}\cup\mathcal{A}$ and by $\mathbf{W}_{\mathcal{A}}$ a random Gaussian vector $\mathcal{N}\left(\mathbf{0}_{p_0+1}, \sigma^4\mathbf{C}_{\mathcal{A}}\right)$. Then $\mathbf{V}_n(\mathbf{u}) \to_d \mathbf{V}(\mathbf{u})$ for every $\mathbf{u}$, with $\mathbf{V}$ defined for $\mathbf{u} = (u_0, \ldots, u_p) \in \mathbb{R}^{p+1}$, by

$$\mathbf{V}(\mathbf{u}) = \begin{cases} \frac{\sigma^2}{2}\mathbf{u}_{\mathcal{A}}^T\mathbf{C}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} + \mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \text{ for } j \notin \mathcal{A} \text{ and} \\ & u_j = u_\ell \text{ for } (j,\ell) \in \mathcal{B}, \\ \infty & \text{otherwise.} \end{cases}$$

Recall the notations introduced just before stating Theorem 3. Any vector $\mathbf{u} \in \mathbb{R}^{p+1}$ such that $u_j = 0$ for all $j \notin \mathcal{A}$ and $u_j = u_\ell$ for all $(j,\ell) \in \mathcal{B}$ has $s_0 + 1$ distinct non-zero values. Denoting by $u_0, u_{j_1}, \ldots, u_{j_{s_0}}$ these values, and setting $\mathbf{u}_{\mathcal{B}} = (u_0, u_{j_1}, \ldots, u_{j_{s_0}})^T$, we have

$$\mathbf{V}(\mathbf{u}) = \begin{cases} \frac{\sigma^2}{2}\mathbf{u}_{\mathcal{B}}^T\mathbf{C}_{\mathcal{B}}\mathbf{u}_{\mathcal{B}} + \mathbf{u}_{\mathcal{B}}^T\mathbf{W}_{\mathcal{B}} & \text{if } u_j = 0 \text{ for } j \notin \mathcal{A} \text{ and} \\ & u_j = u_\ell \text{ for } (j,\ell) \in \mathcal{B}, \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathbf{W}_{\mathcal{B}}$ is a random Gaussian vector $\mathcal{N}\left(\mathbf{0}_{s_0+1}, \sigma^4 \mathbf{C}_{\mathcal{B}}\right)$. Clearly, $\mathbf{V}$ has a unique minimum for $\mathbf{u} \in \mathbb{R}^{p+1}$ such that $u_j = 0$ for all $j \notin \mathcal{A}$ and $u_j = u_\ell$ for all $(j,\ell) \in \mathcal{B}$ and $\mathbf{u}_{\mathcal{B}} = -\mathbf{C}_{\mathcal{B}}^{-1}\mathbf{W}_{\mathcal{B}}/\sigma^2$. Since $\mathbf{V}_n$ is convex we can proceed as in Zou (2006) by using the epi-convergence results of Geyer (1994) to prove the asymptotic normality part.

Similar arguments can be used to get the asymptotic normality part for the logistic case since, as we have shown in the proof of Lemma 1, the limit distribution of $J_{\text{lo}}(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}) - J_{\text{lo}}(\boldsymbol{\beta}^*)$ is that of $(\mathbf{u}^T\mathbf{C}\mathbf{u})/2 + \mathbf{u}^T\mathbf{W}$, with $\mathbf{W} \sim \mathcal{N}(\mathbf{0}_{p+1}, \mathbf{C})$.

Let us now turn our attention to the variable selection consistency. Namely, we have to show that $\forall j \in \mathcal{A}$, $\mathbb{P}(j \in \mathcal{A}_n) \to 1$ and that $\forall j \notin \mathcal{A}$, $\mathbb{P}(j \in \mathcal{A}_n) \to 0$. The first claim is an easy consequence of the previous asymptotic result (see Zou (2006)). To prove the second claim, consider an index $j \notin \mathcal{A}$ and denote by C_j the subset of vertices constituting the connected component of G to which j belongs. Let $C_j^0 = \{\ell \in C_j,\ \beta_\ell^* = 0\}$; clearly, $j \in C_j^0$. Our aim is to prove that $\mathbb{P}(\ell \in \mathcal{A}_n) \to 0$, for all $\ell \in C_j^0$. Towards this aim, observe that the subgradient equations enable to write, for $k = 1, \ldots, p$:

$$\mathbf{x}_k^T(y - \mathbf{Z}\widehat{\boldsymbol{\beta}}^{ad}) = \lambda_n^{(1)} w_k^{(1)} r_k + \lambda_n^{(2)} \Bigg(\sum_{(k,\ell)\in E} w_{k\ell}^{(2)} t_{k\ell} - \sum_{(\ell,k)\in E} w_{k\ell}^{(2)} t_{\ell k} \Bigg)$$

where $r_k = sign(\widehat{\beta}_k^{ad})$ for $\widehat{\beta}_k^{ad} \neq 0$ and s_k is some real number in $[-1,1]$ if $\widehat{\beta}_k^{ad} = 0$; likewise, for any $(k,\ell) \in E$, $t_{k\ell} = sign(\widehat{\beta}_k^{ad} - \widehat{\beta}_\ell^{ad})$ for $\widehat{\beta}_k^{ad} \neq \widehat{\beta}_\ell^{ad}$ and $t_{k\ell}$ is some real number in $[-1,1]$ if $\widehat{\beta}_k^{ad} = \widehat{\beta}_\ell^{ad}$. Introducing the set $\tilde{E} = \{(k,\ell) : (k,\ell) \in E \text{ or } (\ell,k) \in E\}$, and setting $t_{k\ell} = -t_{\ell k}$ for $(\ell,k) \in E$, we have the following more compact form for the subgradient equations:

$$\mathbf{x}_k^T(y - \mathbf{Z}\widehat{\boldsymbol{\beta}}^{ad}) = \lambda_n^{(1)} w_k^{(1)} r_k + \lambda_n^{(2)} \sum_{(k,\ell)\in\tilde{E}} w_{k\ell}^{(2)} t_{k\ell},$$

where, in particular, $t_{k\ell} = sign(\widehat{\beta}_k^{ad} - \widehat{\beta}_\ell^{ad})$ for any $(k,\ell) \in \tilde{E}$ such that $\widehat{\beta}_k^{ad} \neq \widehat{\beta}_\ell^{ad}$. Next, the quantity $M_n(k) := x_k^T(y - \mathbf{Z}\widehat{\beta}^{ad})/\sqrt{n}$ can be decomposed (see the proof of Theorem 2 given by Zou (2006)) as the sum of two variables which both converge in distribution to some normal distribution, so that $M_n(k) = O_{\mathbb{P}}(1)$ as $n \to \infty$, for $\ell = 1, \ldots, p$. Let us note that in the logistic case, property (7) along with assumption **AL1** enables to show that each component of the gradient $\nabla J(\widehat{\boldsymbol{\beta}}^{ad})$ is a $O_{\mathbb{P}}(1)$ as well.

Let us now suppose that there exist some $\ell \in C_j^0$ such that $\widehat{\beta}_\ell^{ad} \neq 0$. In this case, either the set $\mathcal{S}_{neg} = \{\ell \in C_j^0 : \widehat{\beta}_\ell^{ad} < 0\}$ or the set $\mathcal{S}_{pos} = \{\ell \in C_j^0 : \widehat{\beta}_\ell^{ad} > 0\}$ is not empty (or both sets are not empty). If $\mathcal{S}_{neg} \neq \varnothing$, let $b^{\min} = \min_{k\in\mathcal{S}_{neg}} \widehat{\beta}_k^{ad}$. Further denote by L the subset of $\mathcal{S}_{neg}$ of connnected indices ℓ such that $\widehat{\beta}_\ell^{ad} = b^{\min}$. Since $S_{neg} \neq \varnothing$, L has at least

one element. Then, summing up the optimality conditions over the indices in L, we obtain

$$\sum_{k\in L} M_n(k) = \frac{\lambda_n^{(1)}}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \frac{r_k}{|\sqrt{n}\widetilde{\beta}_k|^\gamma} + \frac{\lambda_n^{(2)}}{\sqrt{n}} \sum_{k\in L} \sum_{(k,\ell)\in\tilde{E},\ \beta_\ell^*\neq 0} \frac{t_{k\ell}}{|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma} + \frac{\lambda_n^{(2)}}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \sum_{\substack{(k,\ell)\in\tilde{E} \\ \beta_\ell^*=0\ \&\ \widehat{\beta}_\ell^{ad}>b^{\min}}} \frac{t_{k\ell}}{|\sqrt{n}(\widetilde{\beta}_k - \widetilde{\beta}_\ell)|^\gamma}.$$

Since $L \subset \mathcal{S}_{neg}$, $r_k = -1$, for all $k \in L$, and by definition of L, $t_{k\ell} = -1$ for all ℓ such that $\widehat{\beta}_\ell^{ad} \neq b^{\min}$. Moreover when $\beta_\ell^* \neq 0$ then $\beta_\ell^* \neq \beta_k^*$, and

$$\frac{\lambda_n^{(2)}}{\sqrt{n}} \frac{t_{k\ell}}{|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma} \to_{\mathbb{P}} 0,$$

as n goes to ∞. Since $\lambda_n^{(m)} n^{\gamma/2}/\sqrt{n}$ $(m = 1, 2)$ tends to ∞, the $\sum_{k\in L} M_n(k)$ tends to $-\infty$, which contradicts $M_n(\ell) = O_{\mathbb{P}}(1)$ for all $\ell = 1, \ldots, p$. That leads to $\mathbb{P}(\mathcal{S}_{neg} = \varnothing) \to 1$. If $S_{neg} = \varnothing$, then $S_{pos} \neq \varnothing$, and similar arguments can be used (with maxima instead of minima) to get a contradiction. Putting all this together, we conclude that for all $\ell \in C_j^0$, $\mathbb{P}(\ell \in \mathcal{A}_n) \to 0$.

It remains to show the consistency for the set $\mathcal{B}_n$. As for $\mathcal{A}_n$, we need to prove that $\forall (j,\ell) \notin \mathcal{B}$, $\mathbb{P}((j,\ell) \in \mathcal{B}_n^c) \to 1$ and that $\forall (j,\ell) \in \mathcal{B}$, $\mathbb{P}((j,\ell) \in \mathcal{B}_n^c) \to 0$. Let us prove the first claim. If $(j,\ell) \notin \mathcal{B}$ either ($\beta_j^* = 0$ and/or $\beta_\ell^* = 0$), or ($\beta_j^* \neq 0$, $\beta_\ell^* \neq 0$ and $\beta_j^* \neq \beta_\ell^*$). In the first case, when $j \in \mathcal{A}^c$, we have proved previously that $\mathbb{P}(j \in \mathcal{A}_n^c) \to 1$, so $\mathbb{P}((j,\ell) \in \mathcal{B}_n^c) \to 1$. In the second case, if $j,\ \ell \in \mathcal{A}$, and $(j,\ell) \notin \mathcal{B}$, the asymptotic normality result indicates that $\widehat{\beta}_j^{ad} - \widehat{\beta}_\ell^{ad} \to_{\mathbb{P}} \beta_j^* - \beta_\ell^* \neq 0$; thus $\mathbb{P}((j,\ell) \in \mathcal{B}_n^c) \to 1$. Now let us prove the second claim, using as previously the subgradient equations. Let j be an index of $\mathcal{A}$ such that for some $\ell \in \mathcal{A}$ we have $(j,\ell) \in \mathcal{B}$. Then, for some $1 \leq s(j) \leq s_0$, $j \in \mathcal{A}_{s(j)}$, where $\mathcal{A}_s$ still denotes the set of vertices of the s-th connected component of $G_{\mathcal{B}}$. Suppose that there exists some $\ell \in \mathcal{A}_{s(j)}$ such that $\widehat{\beta}_\ell^{ad} \neq \widehat{\beta}_j^{ad}$. As previously we define $b^{\min} = \min_{k\in\mathcal{A}_{s(j)}} \widehat{\beta}_k^{ad}$ and L the subset of $\mathcal{A}_{s(j)}$ of connected indices ℓ such that $\widehat{\beta}_\ell^{ad} = b^{\min}$. Then, summing up the optimality conditions over the indices in L, we obtain

$$\sum_{k\in L} M_n(k) = \frac{\lambda_n^{(1)}}{\sqrt{n}} \sum_{k\in L} \frac{r_k}{|\widetilde{\beta}_k|^\gamma} + \frac{\lambda_n^{(2)}}{\sqrt{n}} \sum_{k\in L} \sum_{(k,\ell)\in\tilde{E},\ \beta_\ell^*\neq\beta_k^*} \frac{t_{k\ell}}{|\widetilde{\beta}_k - \widetilde{\beta}_\ell|^\gamma} + \frac{\lambda_n^{(2)}}{\sqrt{n}} n^{\gamma/2} \sum_{k\in L} \sum_{\substack{(k,\ell)\in\tilde{E} \\ \beta_\ell^*=\beta_k^*\ \&\ \widehat{\beta}_\ell^{ad}>b^{\min}}} \frac{t_{k\ell}}{|\sqrt{n}(\widetilde{\beta}_k - \widetilde{\beta}_\ell)|^\gamma}.$$

Since $L \subset \mathcal{A}$, the first sum converges to 0 in probability. Moreover, the second sum also converges to 0 in probability, while the third sum tends to $-\infty$, which contradicts $M_n(\ell) = O_{\mathbb{P}}(1)$ for all $\ell = 1, \ldots, p$. We therefore conclude that $\mathbb{P}((j,\ell) \in \mathcal{B}_n^c) \to 0$, for all $(j,\ell) \in \mathcal{B}$, which completes the proof of Theorem 3.

7.5 Competing methods

The (Adaptive) Group Lasso (see Huang et al. (2012) for a recent review) is particularly appealing when features can be naturally divided into K nonoverlapping groups $\mathcal{G}_k$ of dimension d_k, such that $\sum_k d_k = p$. Any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ can then be written $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{G}_1}, \ldots, \boldsymbol{\beta}_{\mathcal{G}_K})^T$, with $\boldsymbol{\beta}_{\mathcal{G}_k} = (\beta_{\mathcal{G}_k 1}, \ldots, \beta_{\mathcal{G}_k d_k})^T \in \mathbb{R}^{d_k}$, and the Group Lasso penalty writes

$$\text{pen}(\boldsymbol{\beta}, \mathcal{G}) = \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{j=1}^{J} \|\boldsymbol{\beta}_{\mathcal{G}_j}\|,$$

The "vanilla" version of the Group Lasso uses only the λ_2-penalty term above: selection is then performed at the group level which means that if a group is selected, all the variables it is made of are selected. The added ℓ_1-norm penalty allows the Group Lasso to exclude some variables within a group, which is willing in the situations considered in our experiments. We use the `spams` package (originally accompanying Mairal et al. (2010)) to compute this method, with adaptive weights derived from initial ML estimates of the parameters (same weights as the $w_j^{(1)}$ used for Adaptive Generalized Fused estimates).

In the joint modeling context, we also consider the Lasso with interaction. More precisely, and with the notations of Sections 2.3, we first select the first stratum as the reference stratum. Then, for every $c > 1$, we can write

$$\boldsymbol{\beta}_c^* = \boldsymbol{\beta}_1^* + \boldsymbol{\delta}_c^*,$$

where $\boldsymbol{\delta}_c^* \in \mathbb{R}^{p+1}$ measures the heterogeneity between $\boldsymbol{\beta}_c^*$ and $\boldsymbol{\beta}_1^*$. This is equivalent to working under the overall model (ruling every observation $1 \leq i \leq n$, with $n = \sum_c n_c$):

$$Y_i \quad = \quad \boldsymbol{\beta}_1^{*T} \mathbf{Z}_i + \sum_{c>1} \boldsymbol{\delta}_c^{*T} \mathbf{Z}_i \mathbb{I}(\mathcal{C}_i = c) + \epsilon_i.$$

The Adaptive Lasso can then be used to select non-zero elements in estimates of both $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\delta}_c^*$, $c > 1$, by minimizing, for an appropriate $\lambda > 0$,

$$\Phi(\lambda) = \sum_{i=1}^{n} \Big(Y_i - \boldsymbol{\beta}_1^T \mathbf{Z}_i + \sum_{c>1} \boldsymbol{\delta}_c^T \mathbf{Z}_i \mathbb{I}(\mathcal{C}_i = c) \Big)^2 + \lambda \Big(\sum_{j=1}^{p} \frac{|\beta_{1j}|}{w_j} + \sum_{c=2}^{C} \sum_{j=1}^{p} \frac{|\delta_{cj}|}{w_{cj}} \Big).$$

We used the relaxed version of this method.

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24, by the MSTIC project of the Joseph-Fourier University, and by the ABS4NGS ANR project (ANR-11-BINF-0001-06).

References

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin : Springer-Verlag, 2011.

F. Bunea, A. B. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

M. El Anbari and A. Mkhadri. On the adaptive grill estimator with diverging number of parameters. *Communications in Statistics Theory & Methods*, To appear, 2013.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13:342–368, 1985.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

C. J. Geyer. On the Asymptotics of Constrained M-Estimation. *The Annals of Statistics*, 22:1993–2010, 1994.

S. Ghosh. Adaptive elastic net: An improvement of elastic net to achieve oracle properties. Technical report, Department of Mathematical Sciences, Indiana, University-Purdue University, Indianapolis, 2007.

C. Gourieroux and A. Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *J. Econometrics*, 17:83–97, 1981.

M. Hebiri and S. van De Geer. The smooth-lasso and other $\ell_1 + \ell_2$ penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011.

H. Heinzl and M Mittlböck. Pseudo R-square measures for Poisson regression models with over or underdispersion. *Computational Statistics and Data Analysis*, 44:253–271, 2003.

H. Höfling, H. Binder, and M. Schumacher. A coordinate-wise optimization algorithm for the Fused Lasso. *Arxiv preprint arXiv:1011.6409*, 2010.

J. Huang, P. Breheny, and S. Ma. A selective review of group selection in high dimensional models. *Statistical Science*, 27(4):481–499, 2012.

J. Jia and B. Yu. On model consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20:595–612, 2010.

K. Knight and W. Fu. Asymptotics for Lasso-type estimators. *Ann. Stat.*, 28(5):1356–1378, 2000.

S. R. Land and J. H. Friedman. Variable fusion: a new method of adaptive signal regression. Technical report, Manuscript, 1996.

C. Leng, Y. Lin, and G. Wahba. A note on lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

P. McCullagh and J. A. Nelder. *Generalized Linear Models. 2nd ed.* New-York : Chapman & Hall, 1989.

N. Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, pages 374–393, 2007.

J. Qian and J. Jia. On pattern recovery of the fused lasso. *Arxiv preprint arXiv:1211.5194v1*, 2012.

A. Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37 (5B):2922–2952, 2009.

H. Sun and S. Wang. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375, May 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B.*, 67:91–108, 2005.

S. Vaiter, G. Peyré, C. Dossal, and J. Fadili. Robust sparse analysis regularization. *arXiv preprint arXiv:1109.6222*, 2011.

S. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business & Economics Statistics*, 25(3):347–355, 2007.

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2007.

H. Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

H. Zou and H. Zhang. On the adaptive elastic net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009.